DOCTORAL THESIS

Relating an Institutional Proficiency Examination to the CEFR: a case study

Kantarcioglu, Elif

Award date:
2012

Awarding institution:
University of Roehampton

Relating an Institutional Proficiency Examination to the CEFR: a case study

by

ELİF KANTARCIOĞLU (BA, MA)

A thesis submitted in partial fulfillment of the requirements for the degree of PhD

Department of Media, Culture and Language

University of Roehampton

2012

Abstract

The primary aim of this study is to investigate the contributions of the CEFR linking process, as stipulated by the *Manual for Relating Examinations to the Common European Framework of Reference for Languages: learning, teaching, assessment* (Council of Europe, 2003), to the validation argument of a university level English language proficiency examination. It also aims to explore the impact of the linking process on the pre-determined or desired level of the examination under study.

This study uses both qualitative and quantitative methods to address the above areas and is comprised of three phases. Phase 1 explores every stage of the CEFR linking process as they are being carried out through field notes, interviews, questionnaires and statistics in order to investigate how well the Manual suggestions capture aspects of validity and guide users in this respect. In Phase 2, the study focuses on an overall investigation of the process through a questionnaire after all stages of linking, viz. familiarisation, specification, standardisation and empirical validation, have been conducted. Finally, Phase 3 examines the Manual itself and its suggestions with respect to validation through a critical analysis of the Manual, a questionnaire and interviews.

The study showed that the CEFR linking process helps users focus on particularly the context, cognitive and scoring aspects of validity at all stages, but mostly at the standardisation stage of the process. Provided that data are accumulated systematically at different stages, at the end of the linking process, those undertaking a linking study can put forward a complete validation argument for the examination in question. However, the Manual fails to provide a model that guides users in this respect. The

process also highlights areas to be considered, should the users set out to design or modify an existing examination to measure at the set desired standards.

TABLE OF CONTENTS

List of Tables

List of figures

List of Acronyms

ACTFL	American Council for the Teaching of Foreign Languages
ALTE	The Association of Language Testers in Europe
ANOVA	Analysis of Variance
APA	American Psychological Association
AREA	American Educational Research Association
BUSEL	Bilkent University School of English Language
CAE	Certificate of Academic English exam (Cambridge ESOL)
CEFR	Common European Framework of Reference
CLB	Canadian Language Benchmarks
COPE	Certificate of Proficiency in English
CRESST	Center for Research on Evaluation, Standards and Student Testing
CTT	Classical Test Theory
DAF	DIALANG Assessment Framework
DAS	DIALANG Assessment Syllabus
EAP	English for Academic Purposes
EALTA	European Association for Language Testing and Assessment
ECD	Evidence Centered Design
ELP	European Language Portfolio
ENDaF	Europäische Niveaubeschreibungen für Deutsch als Fremdsprache (European Level Descriptions for German as a Foreign Language)
EPT	The English Placement Test
ESL	English as a Second Language
ESOL	English for Speakers of Other Languages

ETS	English Translation Studies (BUSEL)
ETS	English Testing Services
FAE	Faculty of Academic English (BUSEL)
FCE	First Certificate in English exam (Cambridge ESOL)
FME	Finnish Matriculation Examination
GEPT	General English Proficiency Test
GESE	Graded Examinations in Spoken English (Trinity College)
ICC	Intra-Class Correlation
IELTS	International English Language System
ILR	Interagency Language Roundtable
ISE	Integrated Skills in English Examinations (Trinity College)
IRT	Item Response Theory
LAB2	Least Able B2
MFR	Many-Facet Rasch
NCME	National Council on Measurement in Education (US)
OPLM	One Parameter Logistic Model
PFIAT	Pre-Faculty Institutional Achievement Test
SEM	Standard Error of Measurement
TDC	Testing Development Coordinator
TEEP	Test of English for Educational Purposes
TOEFL	Test of English as a Foreign Language
TOEIC	Test of English for International Communication
TWE	Test of Written English
UCLES	University of Cambridge Language Examinations Syndicate

Acknowledgements

For supervising my research and widening my horizons in the field of language testing, I am grateful to Barry O'Sullivan, my principal supervisor. Not only did he provide me with guidance and encouragement but also with valuable opportunities to work with others involved in testing. I would also like to thank him for helping me to present my research at different forums and publish my work, as well as being such a good friend, not just a supervisor.

I would also like to express my deepest gratitude to my co-supervisor, the director of Bilkent University School of English Language, John O'Dwyer who worked with me regularly since the beginning of my PhD studies. The fruitful discussions I had with him, his critical comments and reading of my chapters made me a much better writer as well as helping me gain a better understanding of my area of focus. I would also like to thank him for his patience, encouragement and constant faith in me.

I am also indebted to the Bilkent University School of English Language for offering me the opportunity for the PhD and particularly to John for making this possible. This thesis would not have been realised if it were not for the Bilkent University School of English COPE CEFR linking project members, who worked very hard on the project and at the same time putting up with my constant requests for data. I am particularly grateful to Carole Thomas, Hakan Güven, Efser Civelekoğlu and Ayşe Özmen for tolerating my constant abuse and providing me with their invaluable support.

I would like to once again express my deepest thanks to Carole for, despite her unbeliveable workload, always finding the time to discuss any problems I came across in the course of my work, reading my draft chapters and for finally proofreading my thesis on such a short notice.

Finally, this thesis would not have been possible without the the patience and support of my family, especially my husband Selim who had to be both a father and a mother to our daughter during my studies. I owe my greatest debt to my daughter, Deniz, who was born during my PhD studies and had to spend the first years of her life settling for very little attention from me on my lap while I typed away late into the night.

CHAPTER 1

INTRODUCTION

1.1 Background to the study

The Common European Framework of Reference for Languages: learning, teaching, assessment (CEFR) aspires to 'provide a common basis for the elaboration of languages syllabuses, curriculum guidelines, examinations, textbooks, etc. across Europe" (Council of Europe, 2001: 1). The CEFR is claimed to "differentiate the various dimensions in which language proficiency is described, and provide a series of reference points (levels or steps) by which progression in learning can be calibrated (ibid: 7). These reference points are defined in terms of behavioural scales and the core of the framework consists of "a descriptive scheme representing aspects of communicative competence and language use; and a set of common reference levels" (North & Schneider, 1998: 224) that categorize language proficiency into six common levels ranging from A1, the lowest level, through A2, B1, B2, C1 and C2, the highest of the levels.

One of the aims of the CEFR was "to help partners [of the Council of Europe] to describe the levels of proficiency required by existing standards, tests and examinations in order to facilitate comparisons between different qualifications (Council of Europe, 2001: 21). In order to facilitate comparability of language qualifications, the Council of Europe also published the pilot version of the 'Manual for relating language examinations to the CEFR' (Council of Europe, 2003) henceforth the 'Manual', accompanied by a Reference Supplement (Takala, 2004) on standard setting (Kaftandjieva, 2004), qualitative analysis methods (Banerjee, 2004) and quantitative

analysis tools (Verhelst, 2004a, 2004b, 2004c, 2004d). After a colloquium held in Cambridge in 2007, the final version of the Manual was published in 2009 based on the feedback gathered from various Manual users who attended the colloquium. However, the research described here is based on the pilot version of the Manual (Council of Europe, 2003) as the final version had not been published when the Bilkent University School of English Language (henceforth BUSEL) CEFR linking project started.

The Manual stipulates the recruitment of a panel of judges who will take part in the linking study and proposes four stages, viz. familiarisation, specification, standardisation and empirical validation, to the process of relating, in other words, linking examinations to the CEFR.

The purpose of the first stage, familiarisation, is to train the panel of judges in the CEFR with a particular focus on the CEFR scales prior to the following stages in the linking process. It requires the judges to carry out a number of tasks that involves close analysis of the scales. The Manual also encourages users to repeat familiarisation activities before both the specification and standardisation stages.

The specification stage helps define the content coverage of an examination in terms of the CEFR and also aims to determine whether the test under study has been developed following good practice. The outcome of this stage is a CEFR linkage claim at content level.

A close analysis of real test items for receptive skills and sample performances for productive skills takes place during standardisation. The aim of this stage is to set performance standards and cut scores in relation to the CEFR.

Finally at the empirical validation stage, evidence as to the 'internal validity' and 'external validity' (Council of Europe, 2003) of the examination in question is accumulated. Internal validity is collected so as to provide information regarding the quality of the test; and external validation aims to confirm or reject the linkage claims made at the specification and standardisation stages through the use of test analysis methods such as anchoring to an external examination or a measure of the same construct or ability.

Since the publication of the pilot version of the Manual in 2003, its influence has spread outside Europe (e.g. Canada, Korea, Taiwan, US) and various organisations including examining boards (e.g. CITO, City & Guilds, Trinity College, ETS), governments (e.g. Catalunya, Slovenia, Japan) and schools (Hellenic American University, University of Bergen) have linked their examinations to the CEFR. BUSEL, one such organisation, decided to link its English language proficiency examination, the COPE (Certificate of Proficiency English), to the CEFR at the B2 level. To this end, in 2006, it set up a project involving 15 people from various groups in the school such as managers, testers, curriculum developers, textbook writers, teacher trainers and teachers. This thesis presents an overview of the resulting project in the context of analysing the approach suggested in the Manual in terms of test validation theory and practice.

1.2 Rationale for the study

The significant influence of the CEFR has been acknowledged by specialists in the field of testing (e.g. Alderson, 2002; Little, 2005; Vandergrift, 2006; Papageorgiou, 2007a; 2007b; O'Sullivan & Weir, 2011). However, soon after the publication of the Manual, the CEFR had also been the target of criticisms (Weir, 2005; Alderson, 2007), particularly regarding difficulties in using the CEFR for test comparability (e.g. Fulcher, 2004a; 2004b; Huhta et al, 2002; Little et al, 2002). In addition, in 2003, the Council of Europe invited examination providers to pilot the Manual with the aim of collecting feedback on the linking process. By 2006, 40 organizations from 20 countries had participated in the piloting of the Manual (Martyniuk, 2006).

The experiences of those undertaking a CEFR linking study have been published in three invaluable sources (Alderson (ed), 2004; Figueras & Noijons (eds), 2009; Martyniuk (ed), 2011). These sources provide detailed information on various studies on CEFR linking with important reflections on the benefits and implications of CEFR studies, with the latter two particularly focusing on linking examinations. However, such literature lacks focus on validation. The only exception appears to be the City and Guilds project (O'Sullivan, 2009a) which noted that the approach to validation in the Manual was limited and outdated. However, no in-depth investigation of the impact of the linking process on the validity argument of the examination under study was conducted. Furthermore, whether the linking process helps users bring an examination to pre-determined, in other words desired, standards was not explored. This was of particular concern to institutions such as BUSEL that sets its own standards determining the level of language proficiency deemed adequate for academic study.

Therefore, this research aspired to address this gap in the literature through a case study, examining the linking process and the Manual to seek answers to the following main research questions:

RQ1. Does linking an examination to the CEFR provide a comprehensive validation argument?

RQ2. Is the CEFR linking process equally applicable to tests of reading and writing?

RQ3. To what extent does the CEFR linking process help test providers to establish an appropriate level for a test?

1.3 Context of the study

1.3.1 Bilkent University and the School of English Language

Bilkent University, Ankara, Turkey, was the first private foundation university established in the country. Founded in 1984, it admitted its first students in 1986. It consists of nine faculties, two four-year professional schools, two two-year vocational schools, the School of English Language and six graduate schools, and has over 12,000 students mainly Turkish, although 72 other countries are represented. The university also enjoys student exchange agreements with several universities in the US, Canada, Netherlands, Germany, Italy, Denmark, England, France and New Zealand; therefore, English is important in Bilkent. The profile of the university as well as the fact that English is the medium of instruction calls for high quality language education. The latter is given through compulsory language courses as part of a student's departmental

studies and in the English language preparatory program, all managed through the school of English language.

Bilkent University School of English Language (BUSEL) has approximately 300 staff, both local and international, working in three separate programs: the Preparatory program, Faculty of Academic English (FAE) and English Translation Studies (ETS) programs. Students who on arrival do not yet possess an adequate level of proficiency to cope with the demands of academic study in English are assigned to the largest program in the school, the Preparatory Program. Once there, students have a maximum of two years to meet the minimum language requirements to enter the faculty programs. The examination, produced and administered in BUSEL, measuring these requirements is called the Certificate of Proficiency in English (COPE). Students need to successfully complete the Pre-faculty level if they are 4-year faculty students and Upper Intermediate level if 2-year students, to access this exemption test.

1.3.2 The Certificate of Proficiency in English (COPE)

The test specifications for COPE (BUSEL, 2008: 3) state that the examination "tests the English Language proficiency level of students with different language learning backgrounds and defines the minimum level of proficiency required for students who wish to enter a graduate or undergraduate degree course at Bilkent University". It is high stakes as it decides whether students have a sufficient level of language proficiency to cope with academic study in English. Administered three times an academic year, it is taken by approximately 3000 students in any one year, either in September, January or June. In September, new students to the university, as well as those completing the preparatory program, sit the COPE whereas January and June examinations only test

those who are already in the school, viz. preparatory program, or amnesty students. Amnesty students, i.e. students who have been dismissed at the end of two years, have a legal right to take the COPE three times after they are dismissed.

The School of English Language has experienced a significant growth in full-time teaching staff and student enrolment figures since its early years. In line with this growth, two changes in management took place; one in 1989-1990 and the other in 1992. Doubt about practices and standards, amongst other things was "a contributory factor to the change in management in 1989" (O'Dwyer, 2008: 72). A Freshman Entrance Exam, the first in-house proficiency examination introduced in 1990, attempted to implement standardized testing procedures as part of the new management's actions to address the issue.

The COPE evolved from this 'Freshman Entrance Examination', with consultancy from the University of Cambridge Local Examinations Syndicate (UCLES), and was pitched at the first Certificate in English (FCE) examination level, considered the absolute minimum for university study. UCLES and BUSEL had an agreement in which UCLES supported BUSEL in terms of exam design and moderation, offering consultants and also moderating the administration of COPE until 1997. As part of the agreement, BUSEL accepted FCE as an exemption exam. One of the reasons for external support in designing a proficiency examination was the fact that the COPE "provided the basis for the production of comparative statistics on student achievement over time and gave a benchmark" by which success of the Preparatory Program could be judged (O'Dwyer, 2008: 74). Therefore, the exam had to be designed professionally as it provided a benchmark, which set the standards of assessment quality.

Two years after the inauguration of the COPE, a new management decided to develop a new curriculum and, in 1992, a needs analysis was carried out as part of the process. The new curriculum specification introduced in 1993 emphasised academic skills and aimed to change "the learning behaviours and therefore the language performance of students" (ibid: 82).

The COPE is currently broken down into five papers; reading, writing, listening, speaking and language. However, it initially consisted of three papers; reading, listening and writing. Table 3.1 below outlines the changes that took place in the COPE reading and writing papers after its initial design, as these two papers are the focus of this research. With the introduction of a skills syllabus in 1993, the reading paper was revised in the same year to include contextualised vocabulary sections rather than discrete point items, which were taken out. The inclusion of vocabulary sections raises questions about the construct validity of the reading paper. Although vocabulary is seen as an inseparable part of reading, in fact a sub-skill and tested in a reading paper (Grabe & Stoller, 1997; Qian, 2002; Qian & Schedl, 2004; Cohen & Upton, 2006), information regarding the rationale for the inclusion of items testing lexis in the COPE reading paper was not available as there were no test specifications at that time. The concern regarding construct validity of the reading paper stems from the lack of test specifications, which could have justified the approach to the reading paper, and not from the fact that it included vocabulary sections. As discussed in the review of literature, an examination based on multi-divisibility theory can measure vocabulary separately as part of the reading paper.

The writing paper was separated from the Use of English paper in 1994, which was a sound decision as the writing task was not only marked for accuracy of language but for other subskills such as organisation and coherence, which are unique to the skill of writing. By taking the writing task out of the Use of English paper, the school wanted to emphasise the importance of writing skills to reflect the new curriculum. The word limit was extended and a choice of two topics was offered to test takers. Increasing the word limit can be interpreted as an attempt to improve the context validity of the examination while providing a choice of writing topics appealed to the test taker characteristics of the candidates.

Table 1.1 Major COPE Developments Over the Last Fifteen Years for Reading and Writing

Year	Rationale	Reading	Writing
1990	FCE as the target level	25 discrete point items consisting of grammar and vocabulary 3 short (less than a page) texts with 5 MC items – total 15 items	A letter of 100-120 words – very guided – part of the Use of English paper
1993/4	Introduction of an EAP skills syllabus	3 texts with a total of 25 contextualised grammar and vocabulary items 3 short (less than a page) texts with 5 MC items – total 15 items	An extended writing task of 250 words – a choice of two guided topics
1995			An extended writing task of about 300 words – a choice of two topics
1998	Introduction of additional tasks targeting 2-year vocational students	2 texts of 15 contextualised vocabulary items 5 reading texts with a variety of tasks 1 multiple choice cloze text testing vocabulary and grammar	Task 1 – A short letter or report of about 120 words – guided Task 2 – An essay of about 350 words – a choice of two topics
2004	Omission of tasks due to issues of reliability and construct validity	3 parts consisting of a total of 6 reading texts with 35 MC items	An essay of about 350 words – a choice of two topics

The next major change to the COPE came in September 1998 when a variety of new task types were introduced. Two distinct groups of students formed the test taker profile of the COPE examination; 4-year degree faculty students and 2-year degree vocational school students. The fact that these two groups could become eligible to sit the COPE examination after fulfilling the requirements of two different levels – Upper-Intermediate and Pre-Faculty – raised some questions about the validity of the examination, construct validity in particular. The examination was to cater for two groups of students thus having two separate cut scores but its tasks only catered for 4-

year degree students, which put the 2-year students at a disadvantage. To address this, a number of tasks were introduced for the benefit of the 2-year vocational students and two boundaries were set for the two different groups. The most important change was the inclusion of an additional writing task, which was intended as a communicative task targeting the 2-year degree vocational school students. Prior to the transition period, a separate vocational examination, with lower level expectations than the COPE, was administered to the 2-year student population.

The revisions to the COPE were made under the supervision of an external expert and, according to his external report, the new COPE performed well with an overall improvement in terms of the wider range of skills and content, and increased reliability (Allsop, 1998). However, boundary setting for the 2-year group was troublesome due to the small number of 2-year candidates, but currently not so much an issue as the university has decided to discontinue 2-year schools in favour of 4-year faculties.

A new version of the COPE examination was introduced in 2004 after a 2 year extensive revision process that started in 2002 as the institution started to make use of more modern technologies developed in the field of testing, allowing previously unachievable goals to be realised (O'Sullivan, 2011). For instance, setting up an item bank based on Item Response Theory (IRT) allowed for ensuring parallel versions of a test in terms of level, or the use of many-facet Rasch (MFR) with objectively rated tests such as writing. Not only did MFR make it possible to empirically analyse how a writing scale performs, but it also helped monitor rater performances; raters are provided with individualised feedback and training is delivered regularly to increase the reliability of marking.

Questioning the construct validity of the COPE examination was the most significant aspect of the revision process, resulting in three main outcomes. Firstly, qualitative and quantitative analyses were carried out on the existing task types in each paper in order to identify the most effective ones in terms of their construct. For example, several task types were taken out of the examination as it was difficult to agree what they measured. Secondly, the examination was analysed using the Rasch measurement model based on Item response Theory (Baker, 1997; Hambleton, Swaminathan & Rogers, 1991), and the extended Rasch model for dichotomous data by Linacre (1989). Based on IRT, a COPE item bank is now in place, addressing the issue of equivalence and marking of the writing paper has been monitored through the use of many-faceted Rasch analysis since 2004. An item bank is a pool of items calibrated on levels of difficulty and consists of items with "known and invariant measurement characteristics" (Nakamura, 2001: 3) and ensures the level and quality of items. Thirdly, the production of detailed test specifications based on Weir's validation frameworks (2005a) in 2005 allowed for the development, administration and scoring of the COPE examination to be in line with best testing practice. In its current form, the COPE examination aims to measure students' language proficiency through the direct testing of contextualised language performance in the skills of writing, reading and listening, all reflecting academic English. Recently, a speaking component was added to the COPE examination.

In the early years of the COPE examinations, BUSEL committed to the standards set by external experts; however, maintaining standards was done in a traditional and intuitive way. Classical Test Theory was used to analyse the test and parallel versions were ensured through expert judgment only. After production of new tasks, test writers had discussions on the items and text levels to ensure that the test was at the intended level,

supported by classical item analysis. Since 1998, the standards set through the COPE examination have been maintained internally, i.e. without an external consultant. For instance, it was possible to validate COPE results against scores of students on the Pre-faculty Achievement Test (PFIAT) which used to be administered at the end of the preparatory program exit level prior to the COPE and functioned as a 'pre-COPE qualifying exam' (O'Dwyer, 2008: 231) between 1993 and 1998. The school stopped administering PFIAT as it was perceived as a second exit level test by students and the results of these two tests showed variation in pass rates. Various means such as teacher estimates of each student's success at COPE and future performance of students in faculties were used to monitor the validity of the COPE exam. These systems employed to maintain the standards of the examination might be considered rather subjective for two reasons. Firstly, statistical differences or similarities in terms of level were not established empirically to make solid comparisons between different versions of the COPE examination. Secondly, the teacher estimates might have been influenced by differing student characteristics rather than test characteristics. However, O'Dwyer (2008) provided evidence using Rasch analysis that the writing standards in the COPE examination remained constant between 1990 and 2003.

Since 2004, the standards of the COPE exam has been maintained by retaining the quality of the team of writers (ibid) and through test analysis methods such as detailed item writer guidelines, test specifications and Item Response Theory; the latter is not population dependent, unlike Classical Test Theory. To move further forward and away from traditional standard setting, i.e. local expert knowledge, the BUSEL Directorate decided to embark upon a CEFR linking study to better define the standards set through the COPE examination and to construct a validation argument for it.

1.3.3 The CEFR linking project

The BUSEL senior management assigned the Testing Development Coordinator (TDC), also responsible for overseeing the production and administration of the COPE examination, to set up a COPE CEFR linking project. The researcher was also assigned to the project and was granted a scholarship to do a research-based PhD on the linking process. The TDC and the researcher analysed the Manual, prepared a project framework, then requested the Senior Management of the school to form a group of 15 people to take part in the project.

A major aim of the linking project was to verify that the COPE examination was measuring the standards called for by the stakeholders, using the CEFR Manual as the tool to help realise this. As reported by Thomas and Kantarcıoğlu (forthcoming), going through the process required close scrutiny of the examination and would inevitably contribute to its overall quality in two ways: firstly, by using the CEFR as a means "to reflect on their current practice with a view to situating and co-ordinating their efforts and to ensuring that they meet the real needs of the learners for who they are responsible" (Council of Europe, 2001: 1), and secondly, by requiring the examination providers to demonstrate the internal and external validity of their examination (Council of Europe, 2003). In other words, the CEFR linking process would help provide evidence on the validity of the COPE examination for outsiders.

A second aim of the project was to produce CAN DO statements for COPE scores as part of the linkage to the CEFR. It was decided that production of CAN DO statements would raise all stakeholders' awareness of what a COPE score entailed. A further rationale was in line with the Intergovernmental Symposium held in Rüschlikon,

Switzerland in November 1991 (North, 1992). Here it was recognised that there was a need "to provide a sound basis for the mutual recognition of language qualifications" (Council of Europe, 2003: 5). To this end, it was acknowledged that it would be beneficial for stakeholders if the COPE examination were seen both nationally and internationally as a quality examination at B2 level, which would allow students to take this qualification with them when they apply to other universities or colleges of further education. This fits in with one of the aims of the CEFR which is "to facilitate the mutual recognition of qualifications gained in different learning contexts and will aid European mobility" (Council of Europe, 2001: 1).

1.3.4 Participants of the COPE CEFR linking project

Perhaps the most important stage of a CEFR linking project is the standard-setting stage, which as Kaftandjieva (2004: 1) points out "is at the core of the linkage process". Standard-setting is a judgmental process and the participants "are critical to the success of the endeavour" (Cizek & Bunch, 2007: 49). Thus the background of the judges involved in the process and the quality of their judgments are highly significant in the reliability of the cut score established. The project group was formed of 15 participants from the Preparatory Program; the majority of them were unfamiliar with the CEFR and not necessarily trained in making assessment judgments. However, they represented different groups in the school. Involving different perspectives on the project, rather than merely from the people responsible for writing the COPE examination, reduced the risk of bias. Due to other work commitments, attendance fluctuated as did the composition of the group throughout the project, but a core group of 10 people attended constantly, as suggested in the Reference Supplement (Council of Europe, 2004: 23). The project members included: the Director of the Preparatory Program; the Testing

Development Coordinator; the Head of Curriculum and Testing; The Head of the Textbook Project; 2 Heads of Teaching Units; 4 Curriculum and Testing Level Specialists; 2 Teacher trainers; and, 4 classroom teachers. In order to have an outsider perspective on the examination, external experts were brought into the project at specification and standardisation stages.

1.4 Content overview

This chapter, Chapter 1, describes the background to this thesis and the rationale behind the decision to undertake this study.

Chapter 2 reviews the literature regarding four main areas. It first explores the two skills, reading and writing, that are targeted in this study with a view to understanding the theories behind them and how they are currently assessed. It then clarifies the term 'standards' in the field of language assessment and reviews some of the commonly used benchmarks. It moves on to the topic of validity and validation as they are at the heart of this study as described in section 1.2., and finally presents a critical review of a number of CEFR linking studies which have been carried out to date.

Chapter 3 presents the research design and gives background information to the context of the study, Bilkent University School of English (BUSEL) Preparatory Program, highlighting the need to link its proficiency examination to the CEFR. Chapter 3 also outlines the Manual approach to linking and the approach driving the linking project carried out in BUSEL. This chapter, in addition, presents the research instruments used to investigate the CEFR linking project.

Chapter 4 investigates the CEFR linking process as carried out in BUSEL. Each of the four stages of the process (familiarisation, specification, standardisation and validation) is analysed in relation to aspects of validation theory (based primarily on the most important theory to emerge recently, that of Weir, 2005a) and the implications of the CEFR linking process on the institution.

Chapter 5 reflects on the CEFR linking process as a whole. The results of a questionnaire that required project members to look back on all stages of the CEFR linking process and identify parameters belonging to different aspects of validity that were considered at different stages of the project. The chapter not only reviewed the process in terms of the Manual approach to validation, but also looked at how participants in the project perceived the process to impact on local institutional standards.

Chapter 6 presents a critical review of validation approach implied in the CEFR linking Manual, again using Weir (2005a) as a benchmark theory. It also presents findings regarding the contributions of the Manual's linking process to the validity of the COPE examination and the impact of the revisions made to the suggested approach of the Manual.

Chapter 7 presents a summary of the research findings, outlines the limitations of the study and discusses the contributions of the research to the field of testing. The chapter also presents the implications of the study to validation theory, the Manual and examinations to be linked besides offering areas for further research.

1.5 Summary

This chapter aimed to outline the background to this thesis (Section 1.1) and the rationale for the research presented in it (Section 1.2). It also gave brief information on the context of the study (Section 1.3) and offered an overview of the content of the thesis (Section 1.4). The next chapter reviews the relevant literature on assessing reading and writing as well as validity and validation models.

CHAPTER 2

REVIEW OF LITERATURE

2.1 Introduction

The aim of this chapter is to set the background for this study. Areas that are of relevance to relating examinations to the CEFR are reviewed with a view to formulating the research questions and constructing the research design of this study. In section 2.2 the area of EAP assessment with respect to the skills of reading and writing is explored. Section 2.3 clarifies what standards mean with respect to testing and benchmarking, followed by a critical overview of the benchmarks that are commonly used around the world, in order to demonstrate why BUSEL, as well as many other organizations, prefer the CEFR over other benchmarks for aligning their examinations. The following section (Section 2.4) analyses the issues of test validity and validation by giving a brief historical background with the aim of determining the most suitable validation framework that could be employed in CEFR linking studies. Then, section 2.5 looks at some of the CEFR linking studies related to the field of language testing. Section 2.6 draws some conclusions based on the issues that are brought to the surface throughout the review of literature chapter and presents the research questions.

2.2 EAP assessment

2.2.1 Overview

Several EAP examinations are in place worldwide. Some of them are administered locally such as the Test of English for Educational Purposes (TEEP) administered at the University of Reading while some others, like the International English Language Testing System (IELTS) or the Test of English as a Foreign Language (TOEFL), are

administered worldwide. The differences in their approach to assessment result from their distinct constructs, which represent certain beliefs about what language ability entails.

This research focuses on two academic skills, namely reading and writing. The skills of reading and writing require an understanding of one another because these skills are interlinked; people write to be read and likewise read to understand texts written by others. What follows aims to examine theories related to these skills for their potential impact on test construction. For the purposes of this chapter, the word 'theory' is used as an umbrella term and under this term the word 'model' will be used to refer to approaches that explain processes underlying reading or writing. In addition, the word 'taxonomy' will refer to approaches that categorise sub-skills based on complexity. In examining reading and writing theories, firstly, models of reading (2.2.2.1) and writing (2.2.3.1) are presented. Secondly, difficulties associated with assessing reading (2.2.2.2) and writing (2.2.3.2) are analysed. Then, examples of reading (2.2.2.3) and writing (2.2.3.3) assessment are discussed with a focus on issues regarding validity. Finally, key points arising from the discussion regarding the assessment of reading and writing are summarized (2.2.4).

2.2.2 Assessing L2 reading

2.2.2.1 Models of reading

Assessing L2 reading requires an understanding of reading theory and how the theory relates to L2 reading ability. Urquhart and Weir (1998: 22) define reading as "the process of receiving and interpreting information encoded in language form via the medium of print". However, when currently prevalent models of reading are analysed, it

can be readily observed that reading is not such an easy concept to define. The sub-skills are not observable and need to be inferred from tasks that are believed to require the use of them.

Two classes of reading models are found in the literature: *process* and *componential models*. *Process models,* which focus on describing how words are recognized and kept in the memory or when syntactic processing begins, are bottom-up, top-down, interactive and interactive-compensatory approaches to reading. As the name suggests, bottom-up, also known as text-driven approaches (Flesch, 1955; Gough, 1985; La Berge & Samuels, 1985; Finn, 1990), start with words or even letters and move up to sentence level. Once a sentence is processed as a whole, it receives meaning. Cohen and Upton (2006) have recently defined this type of processing as employing linguistic knowledge to create meaning. Top-down or reader-driven approaches (Gove, 1983; Goodman, 1985; McCorrmick, 1988; Weaver, 1990; Dechant, 1991), on the other hand, consider the reader's expectations as the dominant factor in the processing of a text. These approaches claim that the reader brings hypotheses to the text and the text data either reject or confirm these hypotheses. The interactive approach (Stanovich, 1980; Ruddell & Speaker, 1985; Rumelhart, 1985; Barr, Sadow, & Blachowicz, 1990) suggests that while trying to synthesize a text, a reader receives information simultaneously from several different sources, such as his pragmatic knowledge of the language or strategic competence. Interactive-compensatory approaches (Stanovich, 1984; Schraw, Wade & Kardash, 1993) propose that there can be more than one process taking place simultaneously while reading and that "a weakness in one area of knowledge or skill, say in Orthographic Knowledge, can be compensated for by strength in another area, say Syntactical Knowledge" (Urquhart & Weir, 1989: 45).

Componential models, on the other hand, merely focus on describing the components involved in reading without attempting to explain how these components interact (Urquhart & Weir, 1998). The two-component model suggests that in order for reading to take place, a text needs to be decoded and then understood. In other words, 'word recognition' and 'linguistic comprehension' are the two components (Hoover & Tunmar, 1993; Fries, 1963; Perfetti, 1977). The multi-component models (Coady, 1979; Hoover & Gough, 1990; Bernhardt (1991); Hannan & Daneman, 2001) are more varied in their constituents. The model put forward by Coady (1979), for instance, has 'conceptual ability, process strategies' and 'background knowledge' as the components of reading ability, while Bernhardt's (1991) model comprises 'language', 'literacy' and 'world knowledge'.

Besides developing models of reading, several researchers have attempted to identify reading subskills. Some focused on subskills in general (Davis, 1944, 1968; Thorndike 1971) whereas others have tried to differentiate between comprehension skills and inference skills (Carrol, 1969, 1971; Munby, 1978). Amongst those listed above, some have developed taxonomies to define "cognitive skills that have had considerable influence on those who think about psychological abilities and psycholinguistic processing" (Alderson & Lukmani, 1989: 256). Taxonomies have been developed from the theories and they define the complexity of reading sub-skills such as reading for gist and skimming, and identifying the strategies for reading like framing and hypothesis testing. Following the emergence of general language taxonomies, the earliest and the most well known of which is Bloom's taxonomy of educational objectives (Bloom et al, 1956), taxonomies specific to reading were also developed. For instance, Brown and his colleagues (1994), basing their taxonomy on the top-down reading model, suggest that

reading comprehension starts with the cognitive interactional angle before moving on to metacognitive strategies. Ruddell's taxonomy (Ruddell & Speaker, 1985), on the other hand, ranges from very basic literal questions to transactional ones.

Other theories in relation to reading have also been developed. One of these was proposed by Urquhart (1987) who put forward two concepts regarding reading; comprehensions and interpretations. By 'comprehensions' he means "the different products of the reading process, the results of the different standards which readers set themselves, partly because of their purpose in reading, and partly because of the nature of the text" (1987: 387). He suggests that comprehension is influenced by factors such as background knowledge and academic background, resulting in different readings of the same text by different readers. It is these different readings of the same text he calls 'interpretations'. Interpretations, according to Urquhart (1987: 388), "are not under the control of the reader, which is a major factor separating them from comprehensions".

A further set of theories concern the nature of reading; whether reading is unitary: sub-skills cannot be separated; or multidimensional: sub-skills can be discerned. Some studies carried out with the aim of identifying whether sub-skills of reading were divisible or not concluded that the skill of reading had distinguishable sub-skills (Davis, 1968; Gunthrie & Kirsch, 1987; Weir, Yang & Jin, 2000). Some others, however, claimed the opposite and stated that reliable results could not be reached to prove that the skill of reading was divisible (Thorndike, 1973; Rosenshine, 1980; Schedl, Gordon, Carey & Tang, 1996). Khalifa and Weir (2009) suggest that the reason this dilemma could not be resolved was that the population sampling, data analysis tools and the tasks used affected the results of these studies.

A final theory regarding reading is the influence of L1 on L2 reading ability. After reviewing most of the studies available at the time, Alderson (1984) introduced the concept of threshold level by concluding that L2 readers have to possess a certain level of L2 knowledge before they can transfer their L1 reading abilities to the L2 context. He later argued that knowledge of the second language is a more important factor than first-language reading ability (Alderson, 2000). For Enright at al. (2000), four factors affecting L2 reading need to be investigated; viz. (a) whether reading skills acquired in one language can be applied to another; (b) how L1 and L2 similarities might facilitate L2 processing; (c) how cross-linguistic interactions might affect L2 reading; and (d) the degree to which linguistic knowledge has an impact on L2 reading.

These different models, taxonomies and theories put forward about reading complement each other in different ways. The process models attempt to describe the cognitive processes readers undergo and succeed in capturing a different stage in the reading process while the componential models facilitate the understanding of the different types of knowledge a reader resorts to. The taxonomies, on the other hand, differentiate between low level reading abilities and more complex ones. Other theories regarding different reader interpretations, the divisibility of reading or the influence of L2 on L1 reading ability, are all perspectives which contribute to an understanding of the reading skill.

2.2.2.2 Difficulties associated with assessing reading

The analysis of models of reading and other theories regarding the reading skill point to difficulties existing with the assessment of reading ability, which seem to fall into four broad categories. The first difficulty regards the complexity of reading, in that, as

presented in the preceding section, no single model succeeded in defining it. The second stems from different reader interpretations as argued by Urquhart (1987). The third involves the ongoing dispute over the nature of reading, whether it is unitary or multi-divisible. The last difficulty results from L1 influence on L2 reading performance. These difficulties in assessing reading are explained respectively in the subsequent section.

Mathews (1990: 515) indicates that the knowledge and skills involved in reading "are more numerous and complex than currently prevalent taxonomies allow, and furthermore that various kinds of knowledge and skill interrelate differently in the case of individual readers, texts, purposes, etc". In addition, Enright et. al (2000) believe that there are numerous other variables such as text type, task, topic or affect that influence reading comprehension. All these factors involved make designing reading tests troublesome for test constructors as there is no simple way of addressing them in a single test.

Regarding the problem of different readers interpreting texts differently, Urquhart believes that conventional reading tests should be limited to tests of information linguistically 'committed' to the text as the interpretation of the test writer will be different from that of the students (1987: 406). Urquhart and Weir (1998: 115) state that "it is not possible to incorporate 'interpretations' into testing as it is practiced at the moment" and test writers must focus on 'comprehensions'. They also emphasize that interpretations cannot be separated from testees' performance but they shouldn't formally be taken into account in awarding grades.

A further issue that makes assessing reading competence troublesome for test developers is the dispute on whether the skill of reading is unitary or multi-divisible. Weir and Porter (1994), who reviewed several studies carried out to explore the nature of reading, reported the possible dangers of following one or other of these views wholeheartedly as certain individuals might be disadvantaged due to issues resulting from the construct of a test, which would then raise questions regarding the validity of the test. Carrel and Grabe (2002) also propose that readers both engage in processing at different levels known as different reading sub-skills and employ various strategies to facilitate comprehension. Cohen and Upton (2006; 2007) agree that much of the reading process takes place beyond the control of the reader through skills, while at the same time the purposeful use of strategies is also employed. Therefore, test constructors should be very clear as to what their tests are aiming to measure. It is not which view tests should be based on, but whether the intended construct is actually measured successfully that is of concern in assessing reading. Alderson (2000), for instance, draws attention to test validity in as much as it relates to the interpretation of the correct responses to items and claims that if different test takers responded differently to the same item and got it correct, it would be problematic to determine what that item was actually measuring. While he invites test developers to be precise about what they are testing and to be sure that what their test measures is what they claim to test; he admits that it would be "unduly judgmental to blame test constructors for poor tests when theory itself is divided (ibid: 111).

A final difficulty regarding the assessment of reading is the idea of a threshold level for L2 readers proposed by Alderson (1984). This idea should force test constructors to scrutinize what they are testing at lower levels where learners are just beginning to build

knowledge of L2 language. A further challenge is that this threshold level has not been defined and remains a hypothesis.

Test constructors have a challenging job when it comes to assessing reading. For a reading test to have construct validity, all the above difficulties must be addressed; however, this seems almost impossible due to the number of factors involved. The test taker factor, for one, is almost impossible to control in terms of the linguistic and skills background they bring to a test situation. In line with this, it is very hard to predict the impact of test taker performance on text and task types or topics. What kind of cognitive processing takes place in a test taker's mind while performing a reading task can easily show variance from one person to another. This is a vital area that test constructors need a lot more information about, information that, to date, theory has failed to provide.

2.2.2.3 Practice in assessing reading

This section briefly looks at five EAP examinations to investigate the views of reading they are based on and whether their underlying constructs are identifiable with theories. In other words, Alderson's (2000) concern about what tests actually measure, presented above in section 2.2.2.2, is explored. Although the constructs of some of the examinations discussed here are not specified explicitly in documents accessible to everyone, by analyzing the tests and any studies that have been carried out on them, one can make generalisations.

The Test of English as a Foreign Language (TOEFL) reading paper framework clearly states that vocabulary had a significant role in determining the task and item difficulty of the old TOEFL (Qian & Schedl, 2004). This suggests that, as confirmed by Qian and

Schedl (ibid), the TOEFL 2000 reading paper was constructed on a multi-divisible view of reading, which again suggests that the concern raised by Alderson does not seem to be an issue for the old TOEFL, since prior research conducted revealed that "well-designed measures of depth of vocabulary knowledge receive their due recognition as useful predictors of reading performance" (Qian, 2002: 532).

The reading paper of the Test of English for Educational Purposes (TEEP), an English for Academic Purposes (EAP) examination, was claimed to have been grounded on a taxonomic view of language (Weir, 1994). However, in a study carried out on TEEP, Alderson (1990a;b) indicated that the reading skills test items actually measured, contrary to the claims of its constructors, did not reflect the skills the test takers used while responding to the items. This study, based mainly on teacher judgments, revealed that there was a mismatch between the construct of the test and the skills it actually measured. However, Alderson's approach in the study was later questioned by experts who suggested that use of expert judgment without training would yield unreliable results and reported high levels of judge agreement of test items after training (Bachman et al., 1996; Lumley, 1993).

The developers of International English Language Testing System (IELTS) claimed that they were seeking "to sample candidates' ability to perform a number of tasks" from following instructions to drawing logical inferences (Alderson, 2000: 131), suggesting that the paper was constructed on a taxonomic view of reading. A study of the cognitive processing in IELTS showed that the reading test measured both expeditious and careful reading, requiring different response strategies needed in a university context (Weir, et al., 2008). Another construct validation study of the IELTS academic reading test

revealed that the test did reflect a taxonomic view and that it captured the academic domain in terms of 'level of engagement' and 'type of engagement' (Moore, et al, 2008). Furthermore, areas to be refined to better reflect the academic domain were identified and were made by the researchers.

A study conducted on the reading section of the University of Melbourne ESL test involved teacher judgments of sub-skills tested by each item and IRT analysis to investigate whether the sub-skills related to particular items in fact fell into the same difficulty scale. The test was based on a multi-divisible view of reading and the study confirmed this claim. Based on the results of the study, Lumley (1993: 230) states that reading sub-skills represent a useful construct that test writers can work with, although he emphasises that the results of the study do "not suggest that these skills 'exist' in any tangible way".

Although not an academic examination, the First Certificate in English (FCE) is widely accepted as an exemption exam to universities in Turkey as well as some other universities outside Turkey. It is a significant example of a test reflecting real life reading. Weir and Khalifa (2008a) have developed a model for reading based on work done in the field of cognitive psychology. The model sets out to identify key elements of the cognitive processing engaged in by readers in real life tasks. Applying this model to Cambridge ESOL Main Suite Examinations, they found that FCE closely reflected the cognitive processing involved in real life tasks (2008b).

2.2.2.4 Conclusions of assessing reading

As we have seen in the above sections, reading is a complicated skill which has traits or sub-skills that are quite difficult to observe. The lack of any clear theoretical model of reading forces test developers to make choices based on components, sub-skills or taxonomies of reading in the main. In addition, the fact that each individual might process information in a test in a different way puts a strain on the assessment of reading. Therefore, test constructors turn to tasks and what tasks require test takers to do with the aim of capturing specific reading sub-skills represented by these tasks.

2.2.3 Assessing L2 writing

The second academic language skill focused on in this research is writing. This section examines the current thinking on writing and the assessment of writing.

2.2.3.1 Models of writing

The assessment of writing is founded on theories of writing that appear to lie within two perspectives: the cognitive perspective and the sociocultural perspective. The cognitive approach to writing attempts to define what happens within a writer's mind during the act of writing. Hayes and Flower (1980) developed one of the most influential cognitive models of writing in which they describe the writing process as having three parts: task environment; writer's long term memory; and monitoring. Task environment includes the writing task and the text to be produced. Writer's long term memory and monitoring are related to what the writer knows about the topic, audience and planning. These two parts feed into the act of writing, which involves planning what to write, writing it up and then editing it. Hayes (1996) developed the 1980 model by narrowing it down to two main parts and further defining the task environment to include the social and the

physical environment. Furthermore, he integrates motivation or affect and cognitive processes with long-term memory and working memory. Hayes also emphasizes the link between reading and writing in his model and discusses the types of reading that are essential to writing, which are reading to evaluate, reading source texts and reading instructions.

Another influential cognitive model of writing is proposed by Bereiter and Scardamalia (1987) who make a distinction between knowledge telling and knowledge transforming in writing. Knowledge telling, requiring very little planning or revision, is considered natural writing, whereas knowledge transforming requires the use of "writing to create new knowledge" (ibid: 33). Grabe and Kaplan (1996) see this two-model process as an explanation for the differences between skilled and unskilled writers and why writing tasks differ in difficulty even for skilled writers. Similar to the link between reading in L1 and reading in L2, even though strategies used in L1 writing can be transferred to L2 writing experience, L2 writing ability can be hindered by any lack of L2 knowledge (Grabe & Kaplan, 1996; Weigle, 2002).

It has been suggested that cognitive models of writing consider writing as a problem solving task and that skilful writing entails sophisticated problem solving (Deane, et. al., 2008). Expert writers set content and rhetorical goals requiring problem solving whereas inexperienced writers generate one idea that prompts the next one (ibid). The approach taken by the skilled writers is as defined by knowledge transforming and the one adapted by novice writers is explained by knowledge telling as explained above.

Sociocultural approaches to writing propose that the cognitive skills required in writing are "socially situated and take place in social contexts that encourage and support particular types of thinking". (Deane, et al., 2008: 13). Some sociocultural studies in writing have revealed that writing sub-skills exist in a social context and "the institutions and practices associated with them" (Bazerman, 1988; Kamberelis, 1999; Bazerman & Prior, 2005). In other words, the practices in the actual community play a significant role in the type of writing tasks and their structure. For instance, academic writing conventions are determined by academic practices and may show variations depending on the field of study.

Bachman (1990) also argues that language competence comprises pragmatic, organizational and sociolinguistic competences, which take into account all the areas highlighted in the models of the writing process briefly outlined above. Models of language adopting a communicative approach such as those of Hymes (1972), Canale and Swain (1980), Canale (1983), Bachman (1990), Bachman and Palmer (1996) led in the early 1980s to the emergence of communicative language assessment of writing, which is nowadays embraced by many examining boards such as Cambridge ESOL (Khalifa & Weir, 2009).

2.2.3.2 Writing marking procedures

The direct assessment of writing requires marking, where a rating scale or a set of writing criteria is used by a number of raters. Weigle (2002) argued that marking procedures "are critical because the score is ultimately what will be used in making decisions and inferences about writers" (p. 108). A writing score is a result of the interaction among many facets, which include the student's writing ability, the task, the

rating scale and the rater (Hamp-Lyons, 1990; Kenyon, 1992; McNamara, 1996; Weigle, 2002). All these elements involved in a writing score contribute to the writing construct of a test. Therefore, defining the writing criteria and ensuring that raters have a common understanding of the criteria and consistently apply them is crucial to the validity of a writing test.

McNamara (1996) suggests that a rating scale makes implicit or explicit references to the writing construct, in other words the theoretical basis of the required ability and knowledge, upon which the test is grounded. The writing test construct is reflected in the criteria through a number of descriptors defining levels of writing ability and each level of ability is described as a band or score on the criteria. For example, if task fulfilment is an important aspect of the construct, it should be captured in all bands of a rating scale. However, Shaw and Weir (2007) argue that a rating scale on its own is not sufficient in capturing the levels of writing ability "in a way that examiners could reliably and consistently apply" (p. 147), and that exemplar scripts are required to communicate the desired levels of ability reflected in each band or a score on a rating scale.

The exemplar scripts representing scores or bands on a rating scale are used in coordination or standardisation sessions held to train the raters. These sessions aim to ensure that all raters understand the descriptors in the criteria in the same way and apply them consistently. The importance in using exemplar scripts in rater training rather than criteria alone is because "if assessment criteria were separated from students' work they could be interpreted as appropriate for many different levels of achievement" (Wolf, 1995: 76).

Besides rater training, post hoc rating analysis also contributes to the validity of a writing assessment. Intra-rater and inter-rater reliability analyses are two upmost important ways of investigating the rating procedure. "Intra-rater reliability refers to the tendency of a rater to give the same score to the same script on different occasions, while inter-rater reliability refers to the tendency of different raters to give the same scores to the same scripts" (Weigle, 2002: 135). Spearman rank-order correlation, Pearson product-moment correlation coefficient and ANOVA analysis are among ways of analysing rater realibility. Many-faceted Rasch analysis, however, is the most comprehensive of all these types of analysis as it provides information on all facets contributing to a writing score, that is, the rater, the task and the student.

2.2.3.3 Difficulties associated with assessing writing

The testing of L2 writing ability is perplexing due to the complexity of the factors involved as no single definition can cover all the situations in which writing is required and all uses of writing (Weigle, 2002). Models of writing, as with reading, have so much to offer to the assessment of writing because of the difficulty of putting the theory behind the models into practice. As stated above in section 2.2.3.1, communicative views of language competence led to the rise of direct testing of writing, raising several issues related to the concepts of reliability and validity. Assessing writing through communicative-like tasks on the one hand increases reliability by narrowing the range of tasks but, on the other hand, negatively impacts on validity as it restricts the interpretations made through a single task (Saville, 2003; Hawkey & Baker, 2004). In other words, the use of only one task allows for sampling from a single domain of writing ability while more tasks, each targeting a different domain, could make a test

more valid. Hawkey and Baker (2004: 126) point to a further problem by arguing that while communicative tasks "attempt to mirror specific authentic activities, they are also expected to offer generalisability to other task performances and extrapolation to other future abilities". Supporting the problem of generalisability, Deane and her colleagues (2008) suggest that each writing mode or genre deploy a different combination of reasoning, text production and social skills and that each task presents a different problem to the writer thus requiring different cognitive strategies.

The difficulty in designing writing tests that reveal generalisable scores are the parameters that need to be encompassed in them. McNamara (1996) distinguishes between theory-related and pragmatic second language performance assessment, which has had a significant impact on assessing writing. The former refers to language tests that are based on models of language knowledge and language performance such as the one proposed by Hymes (1972), which includes the abilities underlying actual instances of communication. Such theory-related tests focus on language use in context. The latter, on the other hand, aims at designing tasks that reflect the target use situation, with a focus on sociolinguistic correctness. In order to increase generalisability of the interpretations made through test scores, constructs, tasks and target situation requirements should be integrated with rigour in content (Weir, 1993; Bachman, 2002). However, bringing all these parameters to fruition in one writing test is rather challenging for the test constructor.

Similar to McNamara's distinction of theory-based and pragmatic second language assessment, Santos (cited in Cumming, 2002) argued that the prevalent ideology informing expectations for students' L2 writing in academic contexts such as content,

organization and accuracy, is either pragmatic or functional. Students are required to fulfil these expectations for examination or coursework at university. Therefore, their writing ability can only be evaluated in a test context where the use of tests in academic contexts is legitimate.

Cumming (2002) himself points to the mismatch between the type of writing solicited in tests like TOEFL and the writing abilities that students actually have to perform for their academic studies. He indicates that "the contexts of assessment do not correspond directly to the contexts of educational practices" as tests require impromptu and mostly relatively short responses to a given task whereas in academic studies, students are mostly expected to do multiple drafting, use sources, be able to use different genres as relevant and produce an extended piece of writing (ibid: 78).

2.2.3.4 Practice in assessing writing

One obvious option open to test developers is to design tests that balance practical and theoretical aspects of the writing skill, as can be observed in several EAP writing tests. For instance, the analysis of TOEFL in terms of its writing construct reveals that it is limited in terms of genre requirements as it aims to measure candidates' ability to write in a specific genre, viz. argumentative. In terms of authenticity, on the other hand, it reflects academic practice where argumentative prose is of utmost importance. However, TOEFL still leans towards testing based on models of language knowledge and performance since there is "an implicit bias towards privileging linguistic accuracy over other aspects of writing such as task fulfilment and development" (Weigle, 2002: 146).

Cambridge ESOL Main Suite Examinations claim to adopt a communicative writing construct with an emphasis on authenticity, as described by Bachman and Palmer (1996), where the learners are assessed on skills required in the target situation (Shaw & Weir, 2007). In ESOL's First Certificate in English (FCE) examination, however, this claim of authenticity, particularly situational authenticity, is tenuous. To achieve situational authenticity test tasks should reflect real life situations and should be familiar and relevant to the candidature. However, an analysis of the FCE shows that it currently requires test takers to write letters, a skill, one could conclude, recently replaced by writing emails, which is a completely different genre. Furthermore, FCE is an exam administered worldwide. It claims to reflect the theory behind writing, in other words, the cognitive processing involved in writing (ibid). However, the claim that letter writing is a relevant task for test takers all around the world appears tenuous. A further issue relates to the use of FCE as an exemption exam for university entrance and study in the medium of English. In general, and particularly for writing, the task types again seem to lack authenticity for this particular use of the test.

The International English Language Testing System (IELTS), another EAP test managed jointly by Cambridge ESOL, British Council and IELTS Australia, samples from a wider domain of writing ability. Two tasks are included in the IELTS writing paper, targeting different domains in its general training module and the academic module. While the former focuses on language for work and professional training, the latter focuses on language for academic study. The fact that IELTS offers two distinct modules strengthens its construct. However, although task fulfilment is seen as important in the IELTS writing component, as in the TOEFL, when the rating criteria

are considered, the score received is primarily based on language competence, while the communicative element, appropriate task fulfilment, has less weighting.

The English Placement Test (EPT) developed by ETS for California State University reflects the sociocultural model of writing. It is claimed that the EPT is a good example of how critical thinking skills have been considered in writing assessment, in that, rather than tasks that require talking about personal experience, the writing prompts require test takers to present more 'issue-based arguments'. It is also claimed that the writing marking criteria of this test emphasises reasoning skills rather than fluency (Deane, et al., 2008).

2.2.3.5 Conclusions on assessing writing

Assessing writing is a relatively easier skill to define and test than reading as writing is a productive skill resulting in an observable outcome. However, two distinct issues pose threats to the construct validity of some writing examinations. One of them is the issue of making assumptions regarding a test taker's writing ability based on a single task. The second one, and perhaps the more important one, is giving emphasis to linguistic accuracy rather than the sub-skills of writing. These two issues are in line with the two key variables of performance assessment highlighted by Alderson (2005); the task variable and the rating criteria variable.

2.2.4 Summary

Sections 2.2.2 and 2.2.3 have looked at the skills of reading and writing in different ways: theory, assessment and practice. References to a number of examinations were made to demonstrate how the theory is put into practice. The table below (Table 2.1)

summarises the practice of testing and shows that reading tests are mostly designed on a multi-divisible view of reading; and writing tests mostly aim to test communicative language ability. Studies conducted on these tests shed light on what is actually measured in them and helped identify their weaknesses and strengths, as briefly summarised in the 'Comments' column of Table 2.1.

Table 2.1. Overview of how examinations operationalize their constructs

	EXAM	MODEL / CONSTRUCT	COMMENTS
READING	TOEFL 2000	Componential model & multi-divisible view of reading	Task and item difficulty largely based on vocabulary rather than reading skills
	TEEP	Componential model; multi-divisible & taxonomic view of reading	Items targeting different higher order thinking skills and sub-skills
	IELTS	Multi-divisible & taxonomic view of reading	Confidence in what each item is testing and sampling through a number of tasks
	MELBOURNE	Multi-divisible view of reading	Sub-skills useful to work with for test constructors
	FCE	Taxonomic view of reading	Reflecting real life cognitive processing
WRITING	TOEFL	Communicative language ability	Argumentative prose with a focus on linguistic ability
	FCE	Communicative language ability – based on target situation tasks	Samples from a limited domain of writing – letter writing
	IELTS	Communicative language ability – task based assessment of communicative writing construct	Task-based assessment of communicative writing construct
	EPT	Cognitive model of writing – problem solving tasks	Tasks requiring knowledge transforming rather than knowledge telling

2.3 Standards

2.3.1 Overview

Constructing exams without a clear theoretical base is an exigent job, but is not the only challenge facing test designers. Establishing levels or standards is another demanding aspect of testing. The sections that follow explore standards in a number of aspects. First of all, section 2.3.3 attempts to clarify what is meant by 'standards'. Section 2.3.3 defines what benchmarks are and following on from that section 2.3.4 presents a brief review of commonly used benchmarks. Finally, section 2.3.5 draws some conclusions regarding standards.

2.3.2 What are standards?

The word 'standards' in testing is defined by Alderson et al (1995: 236) as "a basis for evaluating test practices" or "a set of guidelines which should be consulted and, as far as possible, heeded in the construction or the evaluation of a test." The concept of standards first emerged with the influence of the *Standards for Educational and Psychological Testing* published in 1985 by the American Educational Research Association (AERA, The American Psychological Association (APA) and the National Council on Measurement in Education (NCME) (Alderson et al, 1995). Several others followed. The *Code of Fair Testing Practices in Education (APA, 2004), the ALTE Code of Practice (1994), ETS Standards for Quality and Fairness (ETS, 2002)* and *EALTA's Guidelines for Good Practice in Language Testing and Assessment (EALTA, 2006)* are the other examples of such standards.

A cursory look at the content of the *1999 Standards for Educational and Psychological Testing* (henceforth the *Standards*) helps to enlighten the issue of standards in testing. The purpose of the *Standards* is three-fold:

- to promote the sound and ethical use of tests
- to provide assessment professionals with guidelines for the evaluation, development and use of testing instruments
- to provide a frame of reference for addressing relevant issues.

The *Standards* mainly deal with the following areas:

- Test construction, evaluation and documentation
- Fairness in testing
- Testing applications

The *Standards* see the issue of validity – one of the main foci of this research - as 'the Grail' of testing (Kaiser & Smith, 2001). They emphasize that testing is pointless without validity. Another area that the *Standards* signify as important entails scales and norms with an emphasis on score comparability that requires linkage and calibration. It is this aspect of testing and assessment, and thus, *Standards* that forms the foundation of the research presented here.

A concept that emerged in line with 'standards', specifically comparability, is standardized testing. A standardized test is a test that "presupposes certain objectives, or criteria, that are held constant across one form of the test to another" (Brown, 2004: 67). The standardized nature of these tests come from the fact that they "specify a set of competences (or standards) for a given domain, and through a process of construct

validation they program a set of tasks that have been designed to measure those competencies" (ibid). A standardized test is designed to yield norm-referenced or criterion-referenced inferences and they are administered, marked and interpreted in a standard way (Popham, 2005). TOEFL and IELTS are among the most commonly known standardized EAP tests. It should be noted though that the ultimate aim of all the sets of standards developed by different organizations is to promote high quality testing and assessment all around the world regardless of the scope of a test. Organizations achieve this by, in general terms, standardizing tasks, marking, administration and levels. Standardizing the first three aspects of testing is done through a set of specifications and is mostly relatively trouble-free. However, the last one, establishing levels, is a demanding task. Recently testing bodies have turned to currently prevalent standards (in Europe these are typically contained in the CEFR) to establish levels for their exams so that the test scores are meaningful to all the stakeholders. They achieve this aim through standard setting, which is defined as "the proper following of a prescribed, rational system of rules or procedures resulting in the assignment of a number to differentiate between two or more states or degrees of performance" (Cizek, 1993: 100). Testing bodies carry out standard setting studies in relation to widely acknowledged benchmarks. What follows aims to give brief information about some of these standards, also called benchmarks.

2.3.3 What are benchmarks?

The perceived need for standards in language learning and assessment has led to the development of "prespecified descriptions of learning outcomes – known, amongst other terms, as standards, benchmarks, competences and attainment targets – as a basis for assessing learners' progress and achievement" (Brindley, 2001: 393). Recently, such

systems have been introduced in a variety of language learning contexts. The Canadian Language Benchmarks (CLB), the Interagency Language Roundtable (ILR) and the Common European Framework for Reference (CEFR) are the most widely known and used sets of language benchmarks. These are discussed briefly below.

2.3.4 Commonly used benchmarks

2.3.4.1 The Canadian Language Benchmark for English as a second language

According to its official website, the Canadian Language Benchmark (CLB) for English as a Second Language instruction is a description of a person's ability to use the English language to accomplish a set of tasks. The Canadian Language Benchmarks were developed with the need to have a common set of standards for measuring language learning among ESL learners. The Canadian Language Benchmarks consist of twelve benchmarks equally distributed in each of the four skills areas.

The aims of the benchmarks are to give:

- information to learners both on what they have learned and what they have yet to learn
- a clear statement of a person's language ability to administrators, teachers, employers, settlement workers and so on
- a set of reference points for teachers to use when assessing a learner's language abilities
- a common basis for assessment of both learners and institutions offering ESL (Centre for Canadian Language Benchmarks, 2005)

The intent of the CLB 2000, the latest version, is to describe communicative language proficiency (See Appendix 2A for CLB 2000). The underlying principle is a belief that language is intended for communication and that the ability to communicate successfully is best described in terms of meaningful task performance within relevant situations and under specific performance conditions (Center for Canadian Language Benchmarks, 2005).

In terms of its construct and the theory behind it, the CLB follows an approach that is recognised and utilised all around the world. However, the limitation of the CLB lies in the context-dependent nature of the descriptors. It attempts to introduce a standardized continuum of competency expressed in a common language that can be used and understood by practitioners across the country. The indicators and descriptors are intended to inform classroom placement, curriculum development and outcomes criteria in Canada (ibid). Therefore, in this respect, the benchmarks are not applicable to be used outside Canada and the Canadian education system and are not intended to cater for language proficiency outside the school curriculum. Vandergrift (2006) agrees with this in that the benchmarks were developed for adult immigrants who learned the language for entry into the Canadian workforce.

2.3.4.2 The Interagency Language Roundtable (ILR) and the American Council for the Teaching of Foreign Languages (ACTFL) proficiency guidelines

The Interagency Language Roundtable scale is the standard scale for language proficiency in the US Federal Service. The ILR scale is, as stated in its official website, a result of the United States Government's effort to define foreign language competence in reaction to the historic inattention to this aspect of its general educational programs. The ILR scale has six base levels descriptions ranging from 0 to 5 for all skills and a

'plus level' description. These 'plus level' grades are assigned when proficiency substantially exceeds that expected for a particular skill level but does not fully meet the criteria for the next level (See Appendix 2B for a summary of the ILR scales). The scale was found to be less suitable for a school context and the lower end of the scale would require revision to cater for this. In the 1980s, the ACTFL developed the lower end of the ILR scale and published Proficiency Guidelines for academic use (See Appendix 2C for the link between ILR scales and ACTFL rating scales).

The studies carried out on the ILR scales are rather restricted in their scope and were mostly carried out on speaking proficiency. Wilson (1999) reviewed a number of ILR based studies that focused on the validity of self-ratings using ILR scales and came to the conclusion, similar to his own research findings on the same area, that ILR-reference self-ratings are valid tools with which to estimate ESL speaking proficiency. On the other hand, the ILR scales have been criticized by Fulcher (2003: 15) for the vagueness of the wording used in the descriptors, and in particular for the fact that "precise definition of terms is avoided."

Another criticism regarding the ILR scales is the references to task types in the descriptors. These task types indicate the tasks the test takers can undertake in the real world. The problem with this is that it limits the kinds of tasks that can be included in the test; as opposed to Bachman's 'ability/interaction' approach which aims to generalize from test scores to real-world situations that may not be modelled in the test tasks (Fulcher, 2003). Vandergrift (2006) also criticized the revised ILR scales, the ACTFL proficiency guidelines, indicating that assumptions were made regarding the stages of second language development which could jeopardize their validity.

2.3.4.3 The Common European Framework of Reference for Languages

The set of benchmarks developed by the Council of Europe is the Common European Framework of Reference for languages: learning, teaching, assessment. In this scheme there are six common reference levels. The following are the aims of CEFR (Council of Europe, 2001:

- to promote and facilitate co-operation among educational institutions in different countries
- to provide a basis for the mutual recognition of language qualifications
- to assist learners, teachers, course designers, examining bodies and educational administrators to situate and co-ordinate their efforts.

The Common European Framework of Reference (CEFR) claims to "provide a common basis for the elaboration of language syllabuses, curriculum guidelines, examinations, textbooks, etc. across Europe" (Council of Europe, 2001: 1). The CEFR "should differentiate the various dimensions in which language proficiency is described, and provide a series of reference points (levels or steps) by which progress in learning can be calibrated" (ibid: 7). These levels or language proficiency scales were developed through a large project in Switzerland (North, 2000; North & Schneider, 1998). (See Appendix 2D for the CEFR global scale).

North (2002) outlines the project as having three main stages. All scales of language proficiency available at the time were put together as the starting point. The descriptors were then undertaken through consultation with a representative group of language teachers so as to determine the most relevant and usable ones. The chosen descriptors,

as the last stage of the project, were used to assess learner performance through questionnaires which were then validated quantitatively using Rasch scaling. The end product was a set of descriptors which could form the basis for common European standards.

However, as these descriptors were developed through the use of questionnaires and self-assessments, there was a need to try them out on actual test results. The first attempt to do so was the Swiss model of the European Language Portfolio (ELP). The ELP was designed to offer learners a procedure which suited "their needs, to empower them to assess and document their language proficiency and intercultural competence, thus enabling and motivating them to plan their further learning" (Lenz, & Schneider, 2002: 68). The Swiss model project followed the same steps as the ones followed in the development of the CEFR descriptors. The end product was a list of descriptors to be used in self-assessment. The only difference between the development of the CEFR and the ELP was that the ELP was subject to a series of piloting studies. The two, in fact, "mutually influenced each other during the development process (ibid: 68). "Even before the European Language Portfolio was officially piloted in Switzerland, it started changing things in directions consistent with the Common European Framework" (ibid: 83).

However, the CEFR is not without its limitations. Weir (2005) criticizes the CEFR from a validity point of view with respect to assessment. He puts forward the following areas of concern:

a) the scales are premised on an incomplete and unevenly applied range of contextual variables and performance conditions (context validity);

b) little account is taken of the nature of cognitive processing at different levels of ability (cognitive validity);

c) activities are seldom related to the quality of actual performance expected to complete them (scoring validity);

d) the wording for some of the descriptors is not consistent or transparent enough in places for the development of tests.

Some other experts also share similar concerns with Weir regarding the use of CEFR in assessment. For instance, Alderson (2007: 661) also points to the problem of theory-based validity by indicating that "there was no theory of comprehension that could be used to identify the mental operations that a reader or a listener has to engage in at the different levels of the CEFR". Similar to Weir, Alderson (ibid) and others (Alderson et. al, 2006) state that the terminology in the CEFR caused problems to the users because of its ambiguous and inconsistent nature. Hulstijn (2007) proposes that the CEFR does not have a strong theoretical basis and that research needs to be conducted on L2 learners rather than only using teacher judgments.

The CEFR has also received numerous positive reactions. Little (2007: 648) indicates that the CEFR had a significant impact on language testing and emphasizes that "the scale seems to offer a ready means not only of indicating the degree of communicative language proficiency confirmed by a particular test or exam, but also of comparing tests and exams with one another, from language to language as well as from country to country". Figueras (2007) states that the interest raised by the CEFR has led to the issue of quality testing receiving more importance. O'Sullivan (2009a: 86) emphasizes that the decision to "embed the CEFR in the development and delivery of its examinations

brought with it a substantial leap forward in the professionalization of the assessment practices of City and Guilds" and that it is the greatest strength of the movement towards benchmarking tests to the CEFR. Regarding the benefits of the CEFR, Papageorgiou (2007a) stated that the undertaking of a CEFR linking project contributed to the quality of the Trinity examinations. In the Slovenian experience of curriculum review project, Pizorn (2009) also points out that attempts to link examinations to the CEFR may lead to improvements of tests as a whole.

2.3.5 Conclusions

Setting standards and establishing levels for language proficiency examinations have been an issue for a long time. When different exams are examined, it can be observed that levels are arbitrarily labelled as A, B or C. This is an issue for all stakeholders; teachers, learners and institutions that make use of such test scores. Therefore, linking examinations to a solid external criterion or benchmark might offer solutions to the dilemma of relating the testing of writing and reading to their theories. Having reviewed the literature on language benchmarks, Vandergrift (2006: 18) concluded that a language framework should possess the following characteristics:

- theoretically grounded (Brindley, 2001; North, 1997)
- empirically validated (Brindley, 1991; North 2000)
- congruent with teachers' perceptions and experiences with language learners (Brindley, 2001)
- transparent and user-friendly (North, 2000; Hudson, 2005)
- context-free but context-relevant (North, 2000)
- comprehensive so that different users can relate their own frameworks and descriptor levels to it (North, 2000)

- flexible and open (North, 2000)
- sufficiently discriminating of levels at the lower end of the framework (Liskin-Gasparro, 1984; North, 2000).

Table 2.2 demonstrates a summary of the strengths and weaknesses of the language frameworks reviewed in this section based on the above criteria as evaluated by Vandergrift (2006: 21). Although CEFR has been criticised for lacking theory and being underspecified (Alderson et al., 2004; Weir, 2005b, Fulcher, 2004a; 2004b; Hulstijin, 2007), Vandergrift claims that the CEFR in fact possesses the characteristics of a good framework, as seen in Table 2.2.

Table 2.2 Summary analysis of the language frameworks reviewed

CHARACTERISTICS	CLB	ILR	ACTFL	CEFR
Theoretically grounded	√	√	√	√
Empirically validated	X	X	X	√
Face validity	√	√	√	√
Transparent and user-friendly	X	X	X	√
Context free and context relevant	X	X	X	√
Comprehensive	X	X	√	√
Flexible and open	X	X	√	√
Sufficiently discriminating of levels at lower end of the framework	√	X	√	X

Adapted from Vandergrift, L. (2006). *Proposal for a Common Framework of Reference for Languages for Canada*. Social Sciences and Humanities Research Council of Canada Heritage.

However, undertaking a linking project (where an empirically supported link is established between a set of standards or benchmarks and a course of study or an examination), no matter how sound the benchmark is, raises other concerns such as how one set of benchmarks – institutional standards – can be linked to another – in this case the CEFR – in a valid manner and what is the role of this linking, that is standard

setting, in the validation argument of an examination under study? In order to answer this question, the following section aspires to investigate areas of validity and validation.

2.4 Test validity and validation

Validity in simple terms is defined as the extent to which a test measures what it is supposed to measure (Henning in Alderson *et al.*, 1995). Messick (in McNamara, 1996: 71) further defines validity as "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores". What is common in these definitions is that validity is not an 'all-or-nothing' matter but a matter of degree (Messick, 1989; Alderson *et al.*, 1995; Weir, 2005a). It is essential to highlight that these definitions of validity also comprise the question 'what for?' As Alderson et al. (1995: 170) put it, 'it is not enough to assert 'this test is valid' unless one can answer the follow-up questions: 'How do you know?' and 'And for what it is valid?' In fact, it is now widely accepted that it is not the test but the inferences drawn from it in relation to the purpose of the test that needs to be validated (Carmines & Zeller, 1979; Messick, 1996; Weir, 2005a).

Looking at the attempts to define validity, a few of which are briefly outlined above, one can clearly see that validity or validation is not easy to pin down. In the past, experts were in a way not able to see the overall picture but only managed to touch a distinct part of the concept and tried to define what they thought validity was. Over the years; however, this has changed into the evolving concept of validity, summarized in the following section 2.4.1. Section 2.4.2 explores validation frameworks with a view to

examining the weaknesses and strengths of each framework and thus propose one to be employed in this research.

2.4.1 Historical overview of validity

Over the years the concept of validity has changed from simply being based on statistical correlations to a complicated concept that involves cognitive, social, contextual or pragmatic as well as psychometric qualities and requires the need to accumulate evidence.

Pre-50s

The concept of validity was very much believed to be all about statistical correlations before the 50s. The test at hand had to be correlated with some criterion, that is, another test claimed to measure the same subject area (Hull, 1928; Bingham, 1937; Guilford, 1946; Gulliksen, 1950). However, the efforts to explore test validity in this period should not be underestimated. Some experts in the field were able to pinpoint some of the indispensable aspects of validity such as content and construct validity and thus brought in a fresh view by introducing the concept of types of validity.

50s and 60s

In the 50s, an empirical orientation to test validity with an emphasis on test use first came into play (Langenfeld, & Crocker, 1994) and then the obligation to accumulate data in an attempt to provide evidence for validity was highlighted (Anastasi, 1954; Thorndike & Hagen, 1955). The boundaries of validity were expanded with the different aspects or types of validity being brought into light. This was particularly reflected in the Standards for Psychological Tests published by the American Psychological

Association (APA) in 1954. However, only four types of validity, namely, content, concurrent, predictive and construct, were considered worth including in those standards. In fact, they were seen as different approaches to validate a test depending on its purpose. In 1955, Cronbach and Meehl, who also worked with the standards committee, combined predictive and concurrent validity into criterion validity, leaving only three types: content, construct and criterion.

70s and 80s

This period saw a major shift in our understanding of validity; experts came to understand that it was not the test itself that had to be validated but "the focus of validation was the inferences and decisions emanating from test scores" (Langenfeld & Crocker, 1994: 152). In the 80s, researchers talked about validity with greater sophistication and suggested a wider range of analytical tools for research (Chapelle, 1999). Chapelle (1999) summarizes the developments regarding validity as follows:

a. the replacement of the three types of validity with a unified view that puts construct at the heart of validity in the 1985 *AREA/APA/NCME standards for educational and psychological testing;*
b. research into the philosophical underpinnings of the validation process;
c. publication of Messick's 1989 paper "Validity", which incorporated test consequences with types of research associated with construct validity.

In this period, a consensus was formed on the meaning of validity as stated in the *Standards 1985,* "the concept (validity) refers to appropriateness, meaningfulness, and usefulness of the specific inferences made from test scores" (cited in Langenfeld, & Crocker, 1994: 152). However, what Messick called attention to in his paper (1989) was

that the social consequences of tests were not incorporated into the notion of validity. Cronbach (1971), again in this period, considered validation as an evaluation argument and suggested that it was the empirical evaluation of the meaning and consequences of measurement

90s and after

After his 1989 paper, Messick's unified view of validity was generally embraced. "This comprehensive view of validity integrates considerations of content, criteria and consequences into a comprehensive framework for empirically testing rational hypotheses about score meaning and utility" (Messick, 1995: 742). According to Weir (2005a), none of the validity types is superior to another and a problem with any one of them raises questions about the soundness of any score interpretation.

The significance of having a validity argument has also been highlighted in this period. Messick (1996: 254) emphasizes the word 'empirical' to show that "the validation process is scientific as well as rhetorical and requires both evidence and argument". Evidence based approaches to validation with a view to putting forward a validity argument have also been developed by Mislevy in the USA and Weir in Europe. These two influential approaches will be returned to below. Meanwhile, Table 2.3 offers an overview of the theories presented above.

Table 2.3. Historical Overview of Validity Theories

	VALIDITY	FOCUS	PROPONENTS
Pre-50s	Statistical correlations	Correlating a test with another	Hull (1928), Bingham (1937), Guilford (1946), Gulliksen (1950)
50s and 60s	Types of validity	Accumulating evidence for content, concurrent, predictive and construct validity	Anastasi (1954), Thorndike & Hagen (1955)
70s and 80s	Unified view of validity	Validating inferences and decisions rather than the test itself.	Langenfeld & Crocker (1994), Chapelle (1999), Messick (1989)
90s and after	Validity argument	Constructing a validity argument through empirical evaluation of the meaning and consequences of measurement.	Messick (1996), Weir (2005)

In the following section an overview of validation frameworks is presented.

2.4.2 Validation frameworks

In line with the evolution of the concept of validity, approaches to validation have also changed resulting in the development of validation frameworks, some of which are briefly presented in Table 2.4 and discussed in this section. Perhaps the first of these was put forth by Cronbach & Meehl (1955). Prior to their influential paper on validity, depending on the purpose of a test, one type of validity (and thus validation method) was chosen. They introduced the idea of creating a 'nomological network' meaning the "interlocking system of laws which constitute a theory" (1955: 10). In order to understand and interpret test results, a network of elements that embodies a test score is generated. Both qualitative and quantitative evidence is accumulated to provide support for each of the components in the network. This model also allows for alterations to the network, meaning in cases where a predicted relation fails to occur, the fault may lie in

the proposed network which in effect leads to redefining the construct. For such new interpretation or network, a fresh body of evidence needs to be collected.

Table 2.4 Validation Frameworks

FRAMEWORKS	FOCUS	SHORTCOMINGS
Cronbach & Meehl (1955)	Nomological network	Intended construct vs. construct proposed as a result of validation
Messick (1989)	Progressive matrix	No guidance on practical side
Shepard (1993)	Hierarchical order of facets of validity	No guidance on practical side
Mislevy (2003)	Evidence Centered Design	Impractical due to numerous models For developing tests rather than analyzing existing ones
Weir (2005)	Socio-cognitive Frameworks	

Cronbach and Meehl's validation framework is significant in its holistic approach to validation; however, it fails in its flexibility to redefine the test construct. A test is designed and developed based on a construct that is determined in advance for a certain purpose. The evidence collected throughout validation is expected to reveal that the test is actually measuring the intended construct. However, according to Cronbach and Meehl's framework, if the outcome of validation demonstrates that the test is measuring a trait other than the intended construct, the construct is redefined based on this evidence, contradicting the notion of test design. In test design, first you decide what you aim to test, that is what traits are important to you and then you design your test to reflect your aim. The ultimate goal is to have a test that measures the intended traits. If, at the validation stage, it is observed that the evidence collected is in conflict with the intended construct, then there are problems with the design of the tasks chosen to reflect the construct, not the construct itself.

Cronbach and Meehl's framework had been embraced for approximately two decades when Messick challenged it with his two-by-two categorization of validity. In his seminal paper, Messick (1989) described validation through a "progressive matrix" "intended to portray validity as a unitary but multifaceted concept (Chapelle, 1999). In his matrix, Messick (1989) conveys the message that the focus of validation should be the relation between evidence and inferences drawn from test results.

In Messick's unitary validation framework (Table 2.5), there are two distinct but interconnected facets which are the source of justification of the testing and the function or outcome of the testing (Messick, 1990). The facet for justification may either have an evidential or a consequential basis. The facet for function, on the other hand, is either test interpretation or test use. The four boxes in his matrix correspond to the four interrelated aspects of the validity question and also present a hierarchical order. At the top of the hierarchical ladder comes the evidential basis of test interpretation which is construct validity since "the evidence and rationales supporting the trustworthiness of score meaning is what is meant by construct validity" (ibid: 23). The evidential basis of test use is of secondary importance and requires evidence for the relevance of the scores to the intended aim and the utility of the scores in the context as well as construct validity. The consequential basis of test interpretation is the appraisal of value implications together with that of the construct itself. Finally, at the bottom is the consequential basis of test use which involves evidential support regarding all the aspects mentioned in the other cells with an additional but new and vital element of validity, that is, social consequences of test interpretation. As seen in the progressive matrix, construct validity is "the integrating force" in this validation framework (ibid: 25). In short, the matrix embodies aspects of validity and assessment about which

evidence pertaining to the hypothesis formed regarding testing outcomes needs to be obtained. However, it does not provide guidance on how this should be done, in that, it is not operationalised.

Table 2.5 Facets of Validity as a Progressive Matrix

	Test Interpretation	Test Use
Evidential Basis	Construct Validity (CV)	CV + Relevance/Utility (R/U)
Consequential Basis	CV + Value Implications (VI)	CV + R/U + VI +Social Consequences

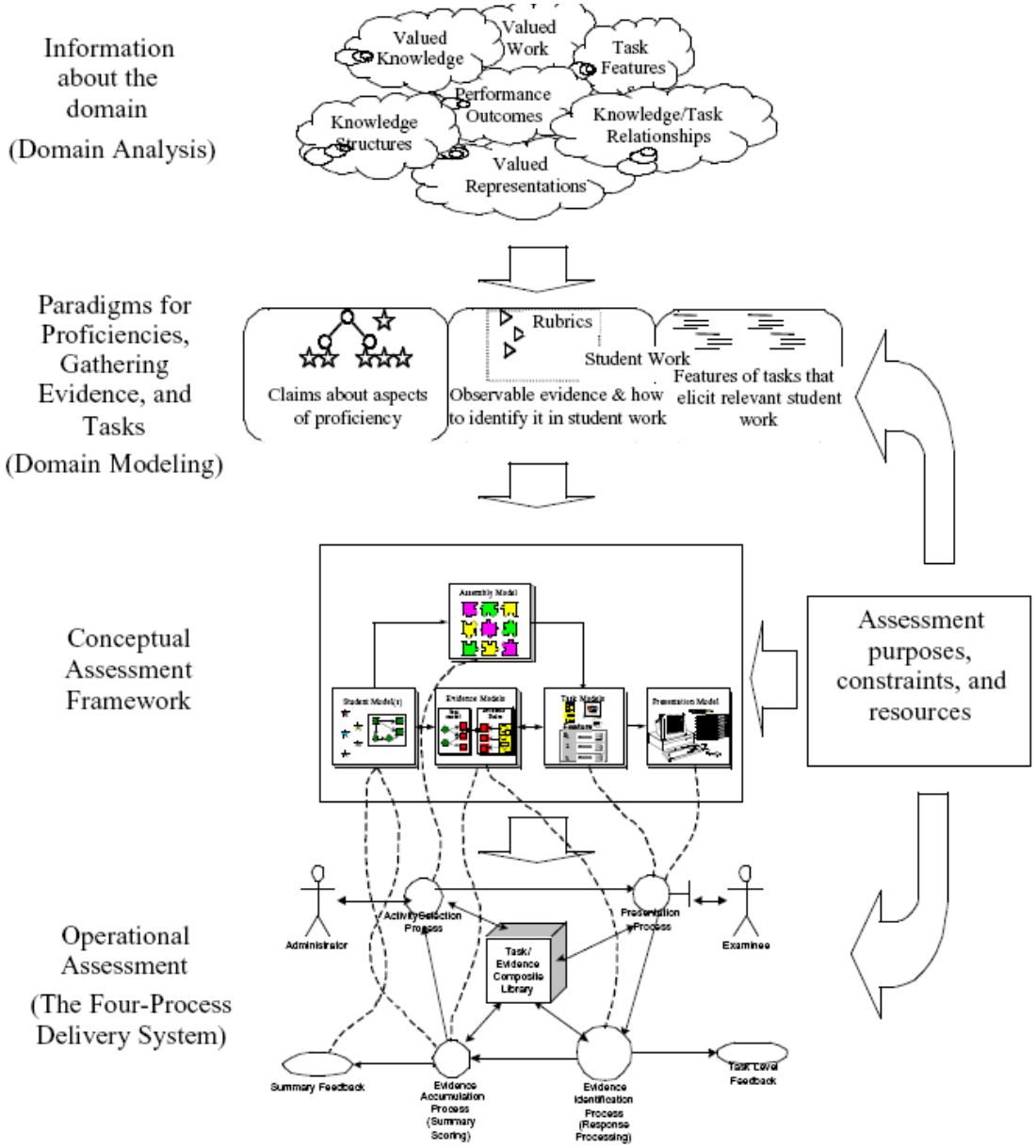
Shepard, however, reformulated Messick's theory and proposed an evaluation argument as the construct validity of test use. Shepard's prime modification was regarding the hierarchical order of the facets that social consequences should come first as questions. This could help identify the validity questions that are essential to support test use (1993). In other words, the argument-based approach to validity organizes "our thinking about important questions and identify priorities" (ibid: 432). The validation framework proposed by Shepard also falls short on the practical side just like Messick's framework. Although Shepard's validation framework gives guidance on identifying important questions and priorities in order to put forward a validation argument, it does not propose how this can be done and how validity evidence to support the argument can be accumulated when it comes to the implementation of the framework.

As seen in Table 2.4, the most recent addition to validation frameworks are those developed by Mislevy and Weir. Both frameworks were constructed with the intention of operationalising the concept of validation. Mislevy et. al (2003: 6) felt the need for a more structured framework "to provide common terminology and design objects that

make the design of an assessment explicit and link the elements of the design to the processes that must be carried out in an operational assessment".

The resulting Evidence Centered Design (ECD) framework developed by Mislevy and his associates (ibid) comprises four main stages: Domain Analysis, Domain Modeling, Conceptual Assessment Framework and The Four-Process Delivery System. In the Domain Analysis stage, the purpose is to collect information about the assessment domain from a variety of sources. The second stage -Domain Modeling-, which according to McNamara (2003) is the most crucial stage, first involves making claims about a student's performance in relation to a test. It then requires considering the type of evidence needed to support the claims made about that student. The final step is about considering the types of tasks that can help obtain the evidence needed. The next stage of design is the Conceptual Assessment Framework, what is known as the blueprint for a test. It provides technical details about the exam such as specifications, operational procedures, statistical tools and rubrics. The last stage, the Four-Process Delivery System, is the stage where all of the above are operationalised and the assessment is delivered (Mislevy et al, 2004). Mislevy (2003: 5) demonstrated the stages of assessment design as shown in Figure 2.1.

Figure 2.1 Stages of assessment design

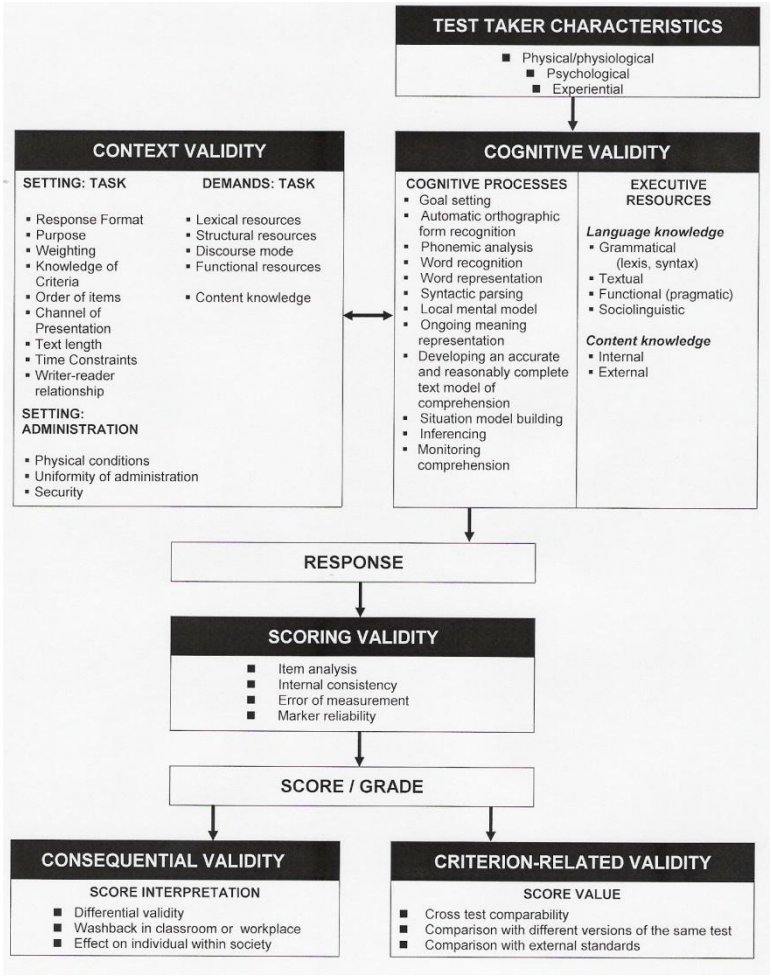


ECD might sound easy to realize, however it has some drawbacks. First of all, even by looking at Figure 2.1, one can understand that despite Mislevy's and his colleagues' efforts to make the process transparent, his proposed framework is still quite theoretical and therefore far from being practical. An example of impracticality is the fact that there are different models to implement the approach and it is suggested that other models should also be developed under this approach to cater for different types of exams and

contexts. This, indeed, suggests that the users should take the theoretical basis of this approach and operationalise validation themselves. There is limited guidance in terms of the operational side of validation. Secondly, ECD is an approach to designing new tests and so its use in validating existing tests is questionable. Although Mislevy states that this framework can be used to analyse existing exams as well as developing new ones, he admits that "it is the latter that should prove more immediately useful" (2003: 62).

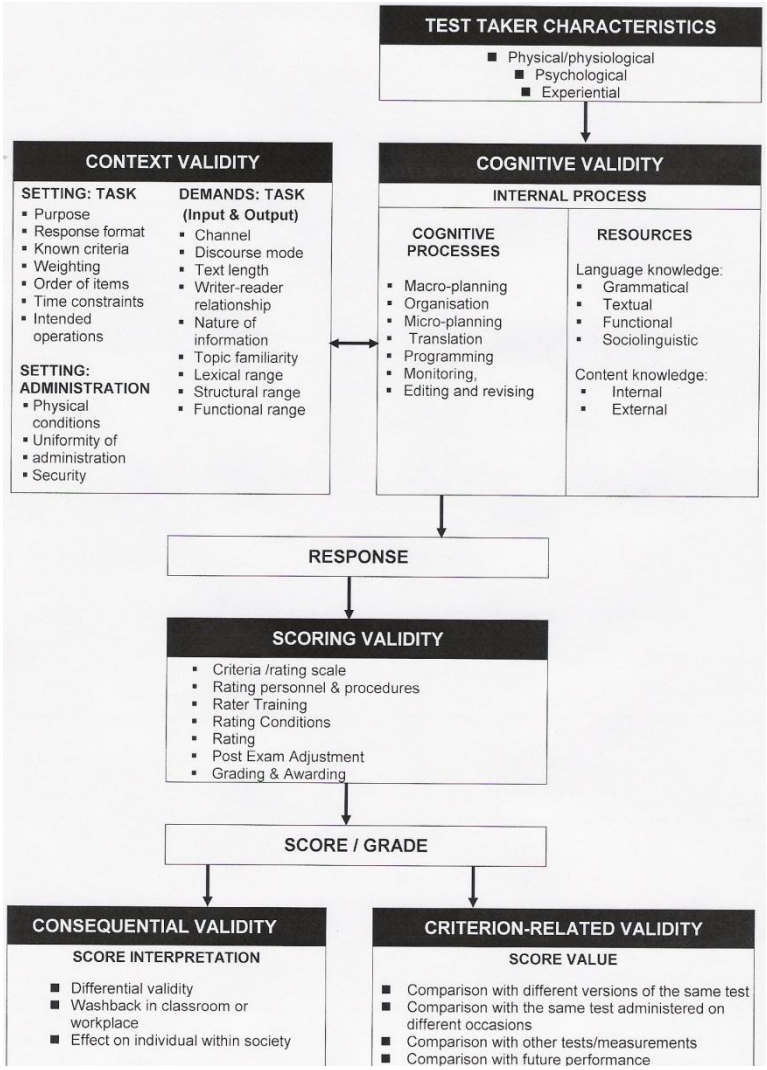
Compared to Mislevy's framework of validation, Weir's approach to validation is much more practical and operationalised in the real sense. Weir presents a more stable framework that can easily be utilized for different exams and in different contexts. He also makes use of commonly used terms and concepts while presenting his validation framework.

Figure 2.2. Weir's Validation Framework (Reading)



The validation framework developed by Weir (2005) takes context validity, theory-based validity, now commonly referred to as cognitive validity, scoring validity, consequential validity and criterion-related validity as the key aspects of validity. He believes these elements need to be addressed by test designers to ensure fairness (ibid). Unlike Mislevy, Weir tackles these aspects of validity separately for each skill, acknowledging the differences between productive and receptive skills in particular, and thus has come up with four socio-cognitive frameworks. Each framework consists of six main parts: test taker, context validity, theory-based validity, scoring validity, consequential validity and criterion-related validity. The frameworks for reading and writing are shown in Figures 2.2 and 2.3 respectively.

Figure 2.3 Weir's Validation Framework (Writing)



Test Taker: The test taker is of great interest to the test developer since the test taker characteristics have a direct impact on the way individuals process the task given.

Context Validity: Under the test taker characteristics comes the physical/physiological, psychological and experiential features. According to Weir's validation framework, task

setting, task demands and administration of the test need to be analysed under context validity (ibid: 44).

Cognitive Validity: The elements of the framework that need to be considered under cognitive validity are the internal processes, namely, executive processes (goal setting, visual recognition, pattern synthesizer) and executive resources (language knowledge and content knowledge). Weir (ibid: 85) clearly emphasizes that it is actually the interaction between context validity, theory-based validity and the scoring criteria that "lies at the heart of construct validity". However, these are tackled separately in his frameworks for ease of description. The term 'cognitive validity' will be used instead of theory-based validity from here onwards as Weir himself preferred this term in a later work (Shaw & Weir, 2007). In addition, Weir and his colleagues later further developed the cognitive validity aspect of his validation model to differentiate contextual parameters from cognitive processing (Shaw and Weir, 2007; Khalifa and Weir, 2009). Both for the reading and writing models, they consider the 'executive resources' in the 2005 model a part of the context validity, i.e. linguistic demands, and cognitive validity involves cognitive processing, where cognitive demands might change depending on the 'cognitive load', i.e. contextual parameters, (Shaw and Weir, 2009) imposed by a reading text, for instance.

Scoring Validity: The scoring validity component of the frameworks is self-explanatory. It focuses on all aspects of the scoring system, from the marking scheme to the selection, training and monitoring of raters (Weir, 2005). Distinguishing factors of assessing each skill is taken into consideration in the scoring systems for different skills.

Criterion-related Validity: Criterion-related and consequential validity are about gathering further evidence on the validity of the test after it is administered and after reliable marking takes place. Criterion-related validity "involves looking for an external criterion beyond the test in question against which it might be measured" (ibid: 207). This includes comparison with another test measuring the same ability, comparison with future performance and comparison with external benchmarks (ibid).

Consequential Validity: Consequential validity, on the other hand, requires the analysis of differential validity (bias), washback and effect on the individual within society (ibid).

In Figures 2.2 and 2.3 Weir clearly demonstrates the link between the types of validity and the kind of evidence that is required for each of them as well as the sequencing of the data collection process. The need for different frameworks for different skills results from the fact that the theory underlying each skill is different, and therefore, requires slightly different types of evidence for validation. It should be kept in mind, however, that the main procedures users have to go through are exactly the same for all of them. That this framework is applicable to all contexts and language exams is a major strength. In addition, Weir's validation framework is the only model that integrates a language theory with a validation theory. Although Bachman's model of Communicative Language Ability is considered to be "the most promising" of the language models, it "has failed to provide a meaningful basis for developing language tests" (O'Sullivan, 2011). It has been criticised for having a psychological basis and being limited in terms of social and cognitive aspects, which critically reduces its suitability for use in language tests (McNamara, 2003; Weir and O'Sullivan, 2011). Bachman's model also fails to adequately account for the modelling of progression in

language ability. The CEFR, on the other hand, offers such a developmental model; however, it does not appear to be supported by a language theory, a situation that has been highlighted by a number of theorists, including Alderson (2007), Fulcher (2004a; 2004b), Huhta et al. (2002), Little (2007) and Weir (2005b). Since it is clear that the validation of linkage claims is essential in any linking project, it became obvious that such a model was needed in order to offer this study the type of theoretical underpinning required. The fact that the only such viable model is that of Weir (2005), which places an understanding of the underlying language model (from both a psychological and social perspective) at the heart of language test development and offers an explanation of how this is connected to validation.

Furthermore, the practicality of Weir's validation framework can be illustrated by a number of studies employing them such as Shaw and Weir, 2006 for writing; Green, (2009) for the Password Test; Khalifa and Weir, (2009) for reading; O'Sullivan (2009a, 2009b, 2009c) for the City and Guilds examinations. These advantages of Weir's framework are the reasons why it was favoured over others and used as one of the primary tools of this research.

2.5 Studies on linking examinations to the CEFR

As mentioned earlier in section 2.3.4.3, since its publication CEFR has faced several reactions, most of which were positive. It has had a considerable impact on curriculum design, self-assessment and particularly language testing. Many institutions have aligned their examinations onto the CEFR. This section analyses some of the CEFR linking studies.

Even though self-assessment could be regarded as a type of assessment, the need to try out the CEFR descriptors on actual test results still remained until the DIALANG project funded by the European Commission under the Socrates program. DIALANG is an on-line diagnostic language assessment system in 14 languages which is based on the six levels of the CEFR scale (Alderson & Huhta, 2005). DIALANG was based on the CEFR in different aspects. "The CEFR concepts of language ability and language use" formed the construct of the DIALANG assessment system (Huhta, A. *et al.,* 2002: 143). The DIALANG Assessment Framework (DAF) and the DIALANG Assessment Specifications (DAS) were drawn from inventories of communicative tasks, themes, activities, types of texts, and language functions in the CEFR. The DIALANG scale was also adapted from the CEFR language proficiency levels. The system works at different levels. Users of DIALANG first take a Vocabulary Size Placement Test, which enables the program to decide at which level of difficulty the exam will be administered. They, then, assess their own language and skill ability through a set of can-do statements. After this initial step of placement procedures, they are presented with a test of the skill and language chosen. In this system, the users get feedback right after they finish the test. In order to develop the system, 14 Assessment Development Teams were formed, one for each language. These teams then wrote over 30,000 items. The trialing was carried out using around 400 learners. The results were analysed using both classical and IRT statistics. Once the piloting was over, the standard setting procedures started. The modified Angoff method (Hambleton & Plake, 1995) formed the first step of the standard setting which was followed by expert judgments. The end product was DIALANG items developed according to a set of specifications based upon the CEFR (ibid).

DIALANG has several important features. It gives the language learners an opportunity to recognize the strengths and weaknesses of their language ability in 14 languages and the feedback learners receive is expressed in terms of CEFR levels. The use of a common yardstick, though an attractive feature, "makes the whole assessment strongly dependent on the quality of this yardstick (Kaftandjieva & Takala, 2002: 106). In the validation study, the DIALANG descriptors used to report the language performance of learners were analysed statistically. According to the results of the study, evidence was collected to claim that the CEFR scales "can be used as a framework for foreign language learning, teaching and assessment" (ibid: 127). Furthermore, when the scales used for reading, listening and writing were compared, it was found out that "the scale for language proficiency in Reading is the best one and that the scale for Writing needs more detailed reconsideration and revision, especially in higher level descriptors" (ibid). The findings of this project raise questions regarding the quality and validity of the CEFR descriptors and scales. Moreover, the project is significant in pointing to the weaknesses of the CEFR while at the same time offering solutions and attracting other users' attention to the problematic areas enabling them to be proactive.

Another study significant in validating the CEFR descriptors was carried out by Brian North (2002) for the University of Basle. Whereas one of the aims of the project was developing a practical self-assessment tool for the University of Basle, the second aim was to validate the CEFR scale of descriptors in one specific context, and illustrate "the way in which the set of CEFR descriptors can be further developed to suit local circumstances" (ibid: 146). The outcome of the project was a self-assessment instrument which had two sections: holistic self-rating and analytic self-rating, which included six skills components. The self-assessment instrument consisted of can-do

statements, some of which were adapted based on the specific context and the rest were from the CEFR scales. Based on the answers given, the total number of points received reflected a CEFR level. Once the instrument was finalized, it was statistically analysed using Rasch which helped determine the range of scores for each level of proficiency. The second step was to correlate the scores on the holistic section and those on the analytic section. According to North, "the relationship between the two self-assessments for the vast majority of learners seems plausible, but with some interesting differences in the distribution" (ibid: 159). The study proved that the difficulty of the CEFR descriptors remained remarkably stable when applied to a different context.

The ENDaF project also had a more global focus than using the CEFR for context-specific purposes and aimed at testing the adaptability of the CEFR. The project aimed to develop consistent level descriptions for German as a foreign language "based upon the Framework's ideas for the levels Breakthrough (A1), Waystage (A2), Threshold (B1) and Vantage (B2)" (Wertenschlag, L. *et al.*, 2002: 184). The results of the project showed that the CEFR has proved to be a useful source for ENDaF in identifying and developing teaching and learning activities. In this respect, the study was significant in demonstrating the adaptability of the CEFR to other languages.

Besides these leading CEFR based studies that have made significant contributions to the field of assessment, several other studies have been carried out around the world in an attempt to bring both curriculum and assessment programs in line with the CEFR. In Catalonia, for example, the CEFR was used as a point of reference both in curriculum design and assessment as well as a reflection tool in different institutional contexts (Figueras & Melcion, 2002). The study mainly involved developing "empirically

validated context-specific proficiency scales related to the CEFR levels" (Generalitat de Catalunya, 2006: 55). These scales would be used to report exam scores. The empirical evidence provided to validate the link of the scales developed to the CEFR was based on correlational analyses i.e. Kendall's coefficient of concordance, Intraclass correlation and correlation of rater judgments as well as alpha reliability. However, these statistical tools may be troublesome in standard setting. Kaftandjieva (Council of Europe, 2004: 23) pointed out that correlational analyses are not appropriate for standard setting purposes as "it is possible to have a perfect correlation of ± 1.00 between two judges with zero-agreement between them about the levels to which descriptors, items, examinees or their performances belong." Additionally, the scale validation was done based on a pair comparison methodology where each descriptor for each level in CEFR and EOI scales that were developed, and for each skill was paired to the descriptor of the same level in the other scale and the rest of the descriptors of both scales. This is a methodology that relies on expert judgment and lacks data from actual learners.

The CEFR was adapted in order to involve learners in the assessment process in primary schools in Ireland. Because "the CEFR descriptors focus mostly on L2 communication outside formal educational contexts, and on the whole imply adolescent and adult language learners/users", the descriptors had to be adapted for the primary school learners' needs (Little, 2005: 328). For the A1, A2, B1, and B2 bands of the CEFR global scale, detailed descriptors were developed for this context. The outcome, which was the ELP for primary school ESL learners, consisted of a passport section, a checklist and a dossier similar to the original Swiss ELP. Once the ELP was prepared, taking the new descriptors for levels of proficiency as the basis, progress, placement and achievement tests were designed. The tasks used in the tests were drawn from the

CEFR, however, the only shortcoming of the study, though a fundamental one, is that the tests were not validated. Therefore, whether the tests served the purpose initially intended or not is still unanswered.

Another CEFR-related study that focused on young learners was the one carried out in Norway. This study is an example of good linking practice in some respects. Similar to the Irish study, the Norwegian one had two main aims; developing portfolio assessment material for lower secondary school pupils compatible with the ELP and a national test of English, which is partially computer-adaptive, to the CEFR (Hasselgreen, 2005). The ELP scale and the can-do statements were adapted for the young learners so as to fulfil the first aim. The second one involved writing items in line with the new can-do statements developed. Items were prepared for each skill and they were then calibrated using the one-parameter logistic model (OPLM), which is an extension of the Rasch model. "The standard setting procedure which followed was largely in line with that outlined in the Council of Europe's Manual and used in DIALANG, and involved both expert judgment and statistical analysis of the items" (ibid). This study is significant in identifying a possible problem with expert judgments. Although the judges were teachers who were familiar with the CEFR and/or experienced, there was little consistency between the judgments and the way items actually performed. Due to this problem, instead of a modified Angoff method, Hofstee's standard setting method was applied, which revealed that the cut off scores were adequate (ibid).

One study that purely focused on mapping test scores onto the CEFR was conducted by ETS for TOEFL, TWE and TOEIC (Tennanbaum, & Wylie, 2005). It is worth noting here that this study was repeated for the new TOEFL (Tennanbaum & Wylie, 2007;

2008). The procedures carried out were similar. The linking study was carried out with two panels of experienced English language teachers and testers around Europe, who recommended B1 and C1 cut scores for the above examinations. The standard setting methods were mainly a modification of the Angoff method and an Examinee Paper Selection Method (ibid). Although the methods used for the purpose of the project were psychometric tools, the results greatly depended on expert judgment, which raises questions of reliability since background information regarding the selection of the expert group is not indicated and no information on why the chosen people are called experts is given in the reports. The group as a whole did not go through training to come to a common understanding of the CEFR levels. This led to disagreements between the judges, which are explained as 'the diversity of relevant professional perspectives' (Tennanbaum & Wylie, 2005: 6) and are ignored. Once the cut offs were set for B2, C1 levels, the plus levels, that is B2+ and C1+, were solely psychometric decisions. The linking, therefore, is only based on the Standardisation stage of the Manual.

The Finnish Matriculation Examination English Test linkage to the CEFR is widely accepted among testers in Europe as a leading example of best practice. The items used in the linkage study were provided to the Council of Europe and offered as benchmarked items to people or organizations that undertake a linking study. However, the study report (Kaftandjieva & Takala, 2002) raises concerns as the CEFR linkage claim has a solely psychometric basis. The intra-judge inconsistency called for aggregation procedures that fit best the empirical difficulty of the items. Item difficulty index was divided into 5 CEFR levels (A2 to C2) with cut score points corresponding to the ends of the confidence intervals of the means. The problem with a purely psychometric linkage to the CEFR is that it entails a number of assumptions. It is first

assumed that items ranging from A2 to C2 actually exist in the examination. Secondly, CEFR levels correspond to almost clear-cut item difficulty values. A final assumption is that the correspondence between CEFR levels and item difficulty is stable for all skills as well as grammar and vocabulary items. Further training of the judges and repeating the standard setting procedures could have helped balance out the expert judgments with psychometric analysis, which could have then led to a more dependable result.

Similar to the case with the Finnish Matriculation Examination, Cambridge ESOL Examinations' items are also recommended as exemplars of level by the Council of Europe to those who undertake a CEFR linking study. It is undeniable that there is a strong relationship between the CEFR and the Cambridge ESOL examinations (North, 2008; Taylor & Jones, 2006; Khalifa & ffrench, 2008). Khalifa and ffrench, in particular, summarizes this relationship with respect to the alignment procedures suggested by the Manual for relating examinations to the CEFR (2003). It is clearly indicated that the Cambridge ESOL main suite exams and the CEFR have a unique relationship in that they have "shared purposes, namely provision of a learning ladder and proficiency framework" and that they have informed "each other's evolution and development" (ibid: 3). In addition, the people involved in item writing and test validation are not only familiar with the CEFR levels but are also trained to write items at the different CEFR levels. Mapping of the examination content onto the CEFR is part of the test development cycle and the items are statistically calibrated and stored in the Local Item Banking System which helps maintain the standards. Internal validation is also given considerable importance but external validation is not carried out (ibid). Although the only missing piece of this puzzle seems to be the external validation, there is a much more serious issue overshadowing the picture and that is the standard setting.

It is true that item writers can be trained to write items at certain levels and these items can be trialled and analysed statistically but what about the issue of 'the intended level versus the true level'? Khalifa and ffrench (ibid) highlight the significance of sustaining the standards but standards need to be set first before they can be maintained. The fact that items have a solid CEFR basis and item writers are trained at producing items at certain levels does not guarantee that these items will function at the levels intended. Standard setting, as stipulated in the Manual, has not taken place, which makes the lack of external validation a concern since some of the procedures suggested in the Manual as part of external validation are in fact ways of standard setting.

Compared to the previous studies, the Trinity College linking study is in a position to make a strong linkage claim due to its comprehensiveness. The Trinity College followed all the stages of the linking process suggested in the Manual while trying to link the ISE and GESE exams. They tried to provide empirical evidence for each stage and took into account the impact of human judgment on the results of the study. The use of a number of psychometric tools; from correlations to Rasch analysis, also demonstrates the solidness of the linkage decisions made (Papageorgiou, 2007a). However, the Trinity study falls short on empirical validation, external validation in particular. The criterion determined as the external criterion is the syllabus that the test is based on. In good testing practice, the test and the institutional syllabus based on CEFR – especially when used as the basis of content specifications for that exam – should have high correlations. However, this high correlation does not necessarily give evidence for the link of the exams in question to the CEFR.

The City and Guilds (O'Sullivan, 2009a; 2009b; 2009c) project to align its examinations to the CEFR was more complete than the Trinity College example. The organisation undertook a major revision study where they first had an expert panel comprised of a small number of people to examine the exam tasks in relation to the CEFR before moving on to the standardisation stage with a larger group. This helped them identify whether their tasks were suitable to be used at B2 level and reflected the B2 level requirements. Their tasks and test specifications were revised based on the feedback received from the critical review expert panel. The standardisation stage was followed by a comprehensive empirical validation stage where issues related to the test taker and context validity were discussed and evidence regarding scoring and criterion-related validity was gathered. The City and Guilds project, however, does not provide evidence as to the quality of the familiarisation and specification stages.

The above review of studies aimed at relating mainly assessment to the CEFR reveals that with the partial exception of the City and Guilds project which adopted a validation theory for the project to present a full validation argument for the examinations, most of the above studies have serious shortcomings with respect to validation. In addition, none of them actually investigated the role of linking and standard setting in the validity argument of an examination. They simply discussed the benefits of linking in general terms as was presented in section 2.3.4.3.

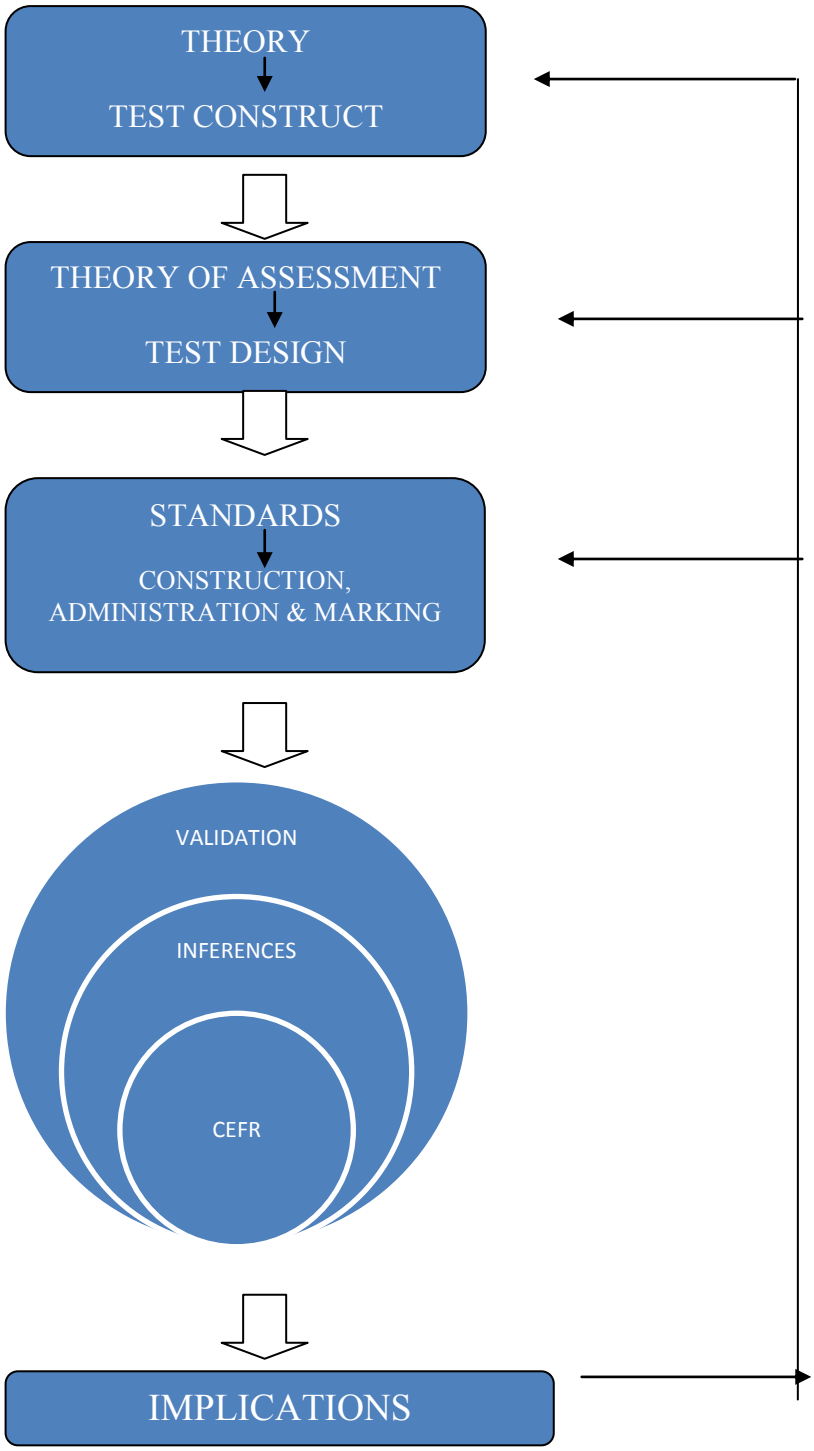
2.6 Summary

In this review of literature, issues related to the theories of reading and writing together with the theory of assessing these skills and how the theory is put into practice; standards and commonly used benchmarks, validity and validation frameworks and

finally studies related to the most widely known and used benchmark – CEFR have been discussed.

The figure below (Figure 2.4) is an attempt to demonstrate the relevancy of and the links between the areas discussed. A clearly described theory of language determines the construct of a language test. The test is then designed in line with a theory of language assessment as well as other facets in Weir's validation frameworks. At the construction stage standards come into play as it is the standards that define the level of the exam. Likewise, standards need to be set for quality exams not only at the construction stage but also at the administration and marking stages to establish that the test functions as intended. Inferences of test results lie at the heart of validation and in this case the inferences are drawn in terms of the CEFR. Any result deriving from the validation process has implications, be they positive or negative, for the test and the organization concerned.

Figure 2.4 An overview of the areas discussed in the literature review



In other words, the aim of any linking project is to make criterion-related (CEFR in this case) inferences about test scores. This is the core of validity and thus validation. It is these inferences that are of utmost importance to institutions. The review of literature

suggests that in terms of validation, the studies undertaken to align exams with the CEFR have several shortcomings. In addition, the purpose of carrying out linking studies raises a further issue of validation. CEFR linking studies are taking place all around the world for reasons such as mobility or sharing a common understanding of what ability at a particular level means. However, no project has attempted to explicitly investigate whether commencing this process adds to the validity of their exam with the exception of the City and Guilds (O'Sullivan, 2009a) project, which suggested that the linking process lead to the professionalization of the assessment practices in the institution. This issue is in fact tied to the implications of such projects and what organizations gain from this experience in terms of the standards they are striving to set. Every language testing body designs examinations with a specific target situation in mind and these examinations are compelled to reflect the level of language competence required in that specific target situation. Universities that produce their own language proficiency examinations for instance, measure academic skills at a language competency level deemed adequate for academic study. The literature reveals that the impact or contribution of CEFR linking studies in setting pre-determined standards has not been investigated.

2.7 Research questions

Literature suggests that conducting a CEFR linking study better defines standards in examinations Pizorn, 2009; Downey & Kollias, 2010; Dávid, 2010; Noijons & Kuijper, 2010; O'Sullivan, 2009a; 2010) through the four stage process suggested by the Manual. BUSEL took a decision to benchmark its exemption examination, COPE, against the CEFR in 2006, well before the studies referenced above. The literature review in Chapter 2 suggests a lack of empirical studies on whether the CEFR linking

process contributes to the validation argument of all aspects of an examination or whether it is more narrowly focused. Furthermore, the implications of the CEFR linking process on the level of an examination have not been investigated. Every proficiency examination sets standards of an intended level they wish to measure. Undertaking a CEFR linking study might reveal that the intended level is not achieved through the examination under study. In such a case, does the linking process help organisations identify areas to be adjusted in order to ensure that the examination reflects the intended level? The current study attempts to fill this gap in knowledge with a particular focus on the skills of reading and writing.

The research questions stemming from the literature review are very broad in focus. To investigate whether linking focuses on all aspects of validity and ensure that evidence is examined for all aspects of validity, sub-questions have been formulated so as to better define the specific focus of each research question. What follows deals with each major research question separately, incorporating sub-questions, the answers to which will contribute to building a body of evidence to provide answers to the major research questions.

3.3.1 Research question 1

Does linking an examination to the CEFR provide a comprehensive validation argument?

This is a key research question which aims to identify what constitutes a solid validation argument. It investigates the degree to which the CEFR linking process, as suggested by the Manual, addresses all aspects of validity. In other words, the first research question

examines whether going through the CEFR linking process focuses on all aspects, or highlights certain aspects of validity at the expense of others.

In order to identify whether this is in fact the case, all aspects of validity have to be examined separately. Therefore, a validation framework, that of Weir (2005a), was chosen, which includes six aspects of validity, each of which is encompassed in the following sub-questions.

1a. To what extent are test taker characteristics taken into consideration during the linking process?

According to Weir's framework (2005a), the test taker lies at the centre of the test development process as test taker characteristics determine almost all aspects of a test such as content, level, topics etc. The evidence required to answer this research question may include whether the CEFR specification forms, used in the second stage of the linking process, ask users to justify the link between the task types or content of a test and the needs or age group of a given candidature. At the standardisation stage, the evidence might be related to whether the task types are suitable for the test takers of a given test.

1b. To what extent does the linking process guide those undertaking a linking study to focus on the context validity of an examination?

Context validity involves design features of an examination in terms of its task demands and administration, and covers parameters such as the purpose of a test, the response format, text length and the content knowledge required in a task. These parameters are essential in any examination and need to be defined in detail at the design stage,

particularly through test specifications. This research looks for evidence of whether parameters of context validity are dealt with in the linking process. For instance, does the linking process encourage users to investigate how well the exam under study achieves its purpose?

1c. To what extent does the linking process focus the attention of those carrying out a linking study on the cognitive aspect of validity of an examination?

The interaction among test taker characteristics, context and cognitive validity lie at the heart of construct validity, therefore, cognitive validity also needs to be clearly defined at the design stage of a test. To explore whether the Manual approach puts the parameters of cognitive validity at the heart of linking, it is necessary to ask where and how in the linking process cognitive validity is tackled. For example, evidence to this effect might be available in the second stage of the linking process if Weir's parameters of cognitive validity, such as language knowledge, appear in the data.

1d. To what extent does the linking process emphasise the importance of the scoring validity of an examination?

Once an examination is designed, it is crucial to ensure that the abilities and competences measured through an examination are strictly reflected and accurately implemented in the marking of that examination. In other words, the design criteria need to be evident in the marking. In order to examine the extent to which the linking process in the Manual ensures scoring validity of an examination, evidence will be collected regarding how and where parameters such as item analysis, inter-rater reliability, or the use of the criteria, are employed in the linking process, if at all.

1e. To what extent does the linking process have an impact on the consequential validity of an examination?

Post-hoc analysis is required to ascertain whether an examination serves its purpose. In terms of consequential validity, Weir (2005a) proposes the analysis of test bias, washback and impact of a test on the society in which it is used. In order to investigate the extent to which the Manual encourages users to focus on this aspect of validity, evidence will be sought in order to identify which, if any, of these parameters are highlighted in the Manual; for example, does the empirical validation stage require users to carry out a study on the backwash of an examination on classroom practices.

1f. To what extent does the linking process have an impact on the criterion-related validity of an examination?

Finally, Weir (2005a) also suggests that it is crucial to compare an examination with its other versions, and an external test measuring the same ability, to ensure the accuracy and stability of its level. Evidence to address this question might come from the empirical validation stage where a number of methods are suggested to users to support the criterion-related validity of an examination.

3.3.2 Research question 2

Is the CEFR linking process equally applicable to tests of reading and writing?

The second research question looks at whether the linking process is similarly applicable to productive and receptive skills, in this case, reading and writing. In other words, does the CEFR process favour one skill type over another? Since this question examines the applicability of the validation approach of the Manual across the board, and whether changes might be necessary to account for the different skills, all aspects of

validity, separately for reading and writing, need to be examined. While sub-questions similar to the first research question are required to address and differentiate the applicability of the CEFR to receptive and productive skills, the evidence required might differ. For instance, does the linking process provide guidance for the empirical validation of a reading test and a writing test in equal terms?

What variations, if any, are there in the Manual's methodology to the validation of productive and receptive language tests in terms of attention to

2a. test taker considerations?

2b. context validity?

2c. cognitive validity?

2d. scoring validity?

2e. consequential validity?

2f. criterion-related validity?

3.3.3 Research question 3

What implications can be drawn from the study for increasing standards by a pre-determined amount?

The final research question is concerned with whether the CEFR linking process offers concrete suggestions to alter the level of an examination if, throughout or by the end of the process, the test does not correspond to the required level the institution undertaking the study desires. The research attempts to collect evidence in the linking process, and in the Manual itself, which guides users on how to address this issue.

3a. How does the linking process contribute to the understanding of the institutional standards set through the examination?

In order to answer this sub-question, first of all, it is necessary to investigate whether the linking process helps institutions better understand the level of the examination under study so that they can make a decision as to how closely the examination reflects the desired level. Evidence contributing to an answer might come from the activities carried out in the process that forces users to analyse their test closely and identify test features that help establish its level. Further evidence might be gleaned from users of the Manual themselves as to how confident they are in applying the CEFR to deal with this challenge.

3b. Can the linking process suggest ways in which an examination could be modified to raise the level of the examination to pre-determined standards?

This question suggests a need for evidence indicating that there are clear pathways that help testers recalibrate their exam to the desired level; evidence might also be suggested within the standard setting process; or, further clues as to how to recalibrate, up or down, in detailed instruction in the Manual or setting intermediate levels, e.g. B1.1 or B1.1.2.

CHAPTER 3

RESEARCH DESIGN

3.1 Introduction

This chapter presents the research design of this study. Firstly, the approach adapted in this study is presented. Secondly, research tools at all phases of the study are presented and justified in the data collection framework. Finally, issues regarding research ethics are discussed.

3.2 Research approach adopted and variables

A case study approach was used to explore the implementation of the CEFR linking process. Thomas (2003: 33) defines a case study as consisting of "an entity and the entity's actions. Frequently, case studies also offer explanations of why the entity acts as it does". The entity or unit of analysis (Yin, 2009) in this case is the project carried out in BUSEL to link its proficiency exam, COPE, to the CEFR. It should be noted that there are degrees of linking, as explored in Chapter 2 section 2.5, and that the project undertaken by BUSEL is a full linking study. The CEFR linking Manual and the procedures suggested there form the content of this research, that is, the material that is researched. BUSEL followed the procedures suggested in the Manual throughout the project. However, although institutions going through the process of linking their examinations to the CEFR are all advised to follow a set of procedures and guidelines determined by an 'external' source, i.e. the Manual, each institution is unique in its interpretation and implementation of those procedures. As will be discussed in detail, particularly in Chapters 4 and 5, local contingencies were reflected in the institutional approach to how the Manual was used. The case study also comprises two sub-units of

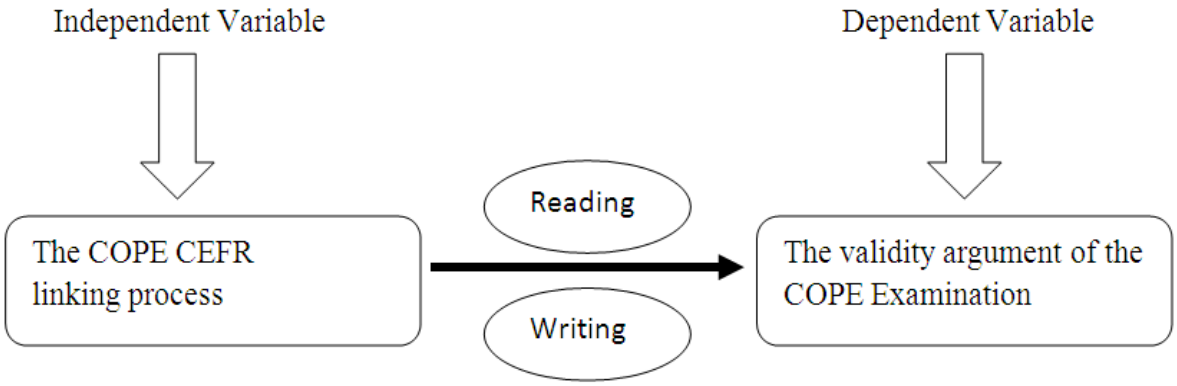
analysis – writing and reading – as it may be fruitful when a case study has comparative elements to enrich its data and findings. The comparative elements result from the fact that writing is a productive skill whereas reading is a receptive one and the distinct nature of these skills call for different standard setting methods and marking procedures.

Stake (1995) proposes three types of case study: intrinsic, instrumental, multiple or collective case studies. In an intrinsic case study, the focus is on the case solely because of its own value. In an instrumental study, the case itself is of secondary importance and the primary focus is on gaining insight into a wider issue through that case study. Finally, in a multiple or collective case study, a number of cases are studied together to gain understanding of a certain phenomena. This particular study encompasses elements of all three types of case study. It is intrinsic in that the project reflects BUSEL's unique approach and implementation of the procedures suggested in the Manual and therefore, understanding the nature of the project as implemented in BUSEL and its implications on the COPE examination is of utmost importance to the institution. It is also instrumental as the aim of the researcher is to come to generalisable findings that are themselves relevant to any external benchmarking study. This study can also be considered a multiple or collective case study since it allows for comparison of the intra-case elements between reading and writing.

The concept of variables is usually associated with quantitative research (Creswell, 2009); however, it seems relevant to all types of research. In this case study as presented in Figure 3.1., the independent variable is the COPE CEFR linking project encompassing the CEFR linking procedures recommended by the Manual and the unique approach adapted in BUSEL in their implementation. The independent variable

directly impacts on the dependent variable, viz. the validity argument of the COPE examination. The impact of the independent variable on aspects of validity as proposed by Weir (2005a), theory-based (more recently referred to as cognitive), context, scoring, criterion-related, and consequential validity, are investigated through the case study. Two embedded elements are under study; viz. reading and writing standards, specifically, the impact of the linking process on setting the standards of the COPE reading and writing papers separately is researched.

Figure 3.1 Case Study Variables



A limitation of case studies lies in the applicability of generalizations or principles drawn from one case to other cases. This becomes problematic when people "are not interested solely in the outcomes of a particular investigation but are interested in how the report of a given case can help them understand other similar people, institutions, or events" (Thomas, 2003: 35). To address this issue in this study, Yin's (2009: 14-124) three principles of data collection for establishing the construct validity and reliability of the case study evidence are embarked upon:

Principle 1 Multiple sources of evidence

Principle 2 Creation of a case-study database

Principle 3 Establishment of a clear chain of evidence

The piece of research uses multiple sources of evidence ranging from field notes and interviews to questionnaires and statistics (Principle 1) as discussed in detail in section 3.7. A database was created, including all documents and session plans used, records of participant judgments, raw data and field notes kept, and statistical analyses conducted throughout the process (Principle 2). Finally, a clear chain of evidence was developed in the design of this study (Principle 3). Data was accumulated at all stages of the CEFR linking process through different tools and at the end of the process to reflect on the process as a whole. The details of how principles 1 and 3 are realized are described in some detail in Section 4.4.

The research is a mixed methods study, involving both quantitative and qualitative data collection and associated analytical methods, and follows Dornyei's principled mixing approach. Dornyei indicates that when different methods are combined in a principled way, "their strengths aggregate, thereby making the sum greater than the parts" (2007: 167). For example, throughout the CEFR linking process, data was collected through field notes and interviews but at the end of the linking process, a questionnaire targeting all aspects of validity, including those that had not come out of the analysis of the field notes and interviews, was administered.

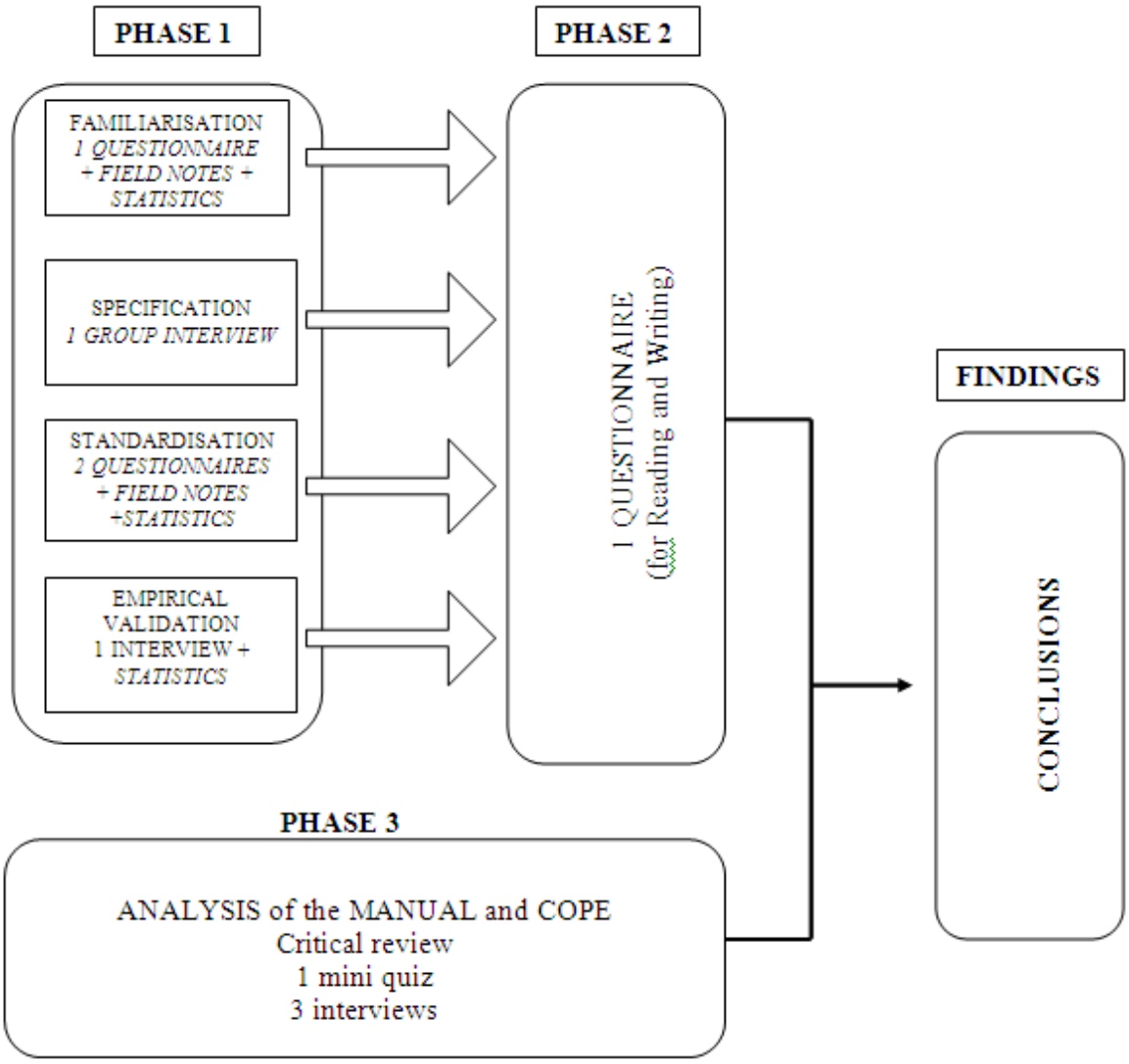
Mixed methods research design involves sequential or concurrent combinations of methods with an emphasis on either the qualitative or quantitative method(s) or both. In this research concurrent combinations of qualitative and quantitative methods are used to broaden the research perspective and validate the resulting hypotheses by

triangulating data using multiple methods, which is one of the benefits of mixed methods research. Triangulation helps in reducing "the risk of bias or chance resulting from a specific method. It also contributes to a better assessment of the generality of the explanations made (Maxwell, 1996). In concurrent designs, two methods – qualitative and quantitative – are used in a separate and parallel manner and the results are integrated when they are interpreted. In Phase 1 of this research, both qualitative and quantitative methods are used in a parallel manner to collect data. In Phase 2, a qualitative method follows a quantitative method in order to gain a deeper understanding of the findings reached as a result of the former tool. In Phase 3, a qualitative and a quantitative tool are used again in a parallel manner but this time to compare the results of the two data collection methods. These phases are explained in detail in section 3.3.

3.3 Data collection framework

In order to research the impact of the linking process on the validation argument of the COPE examination, data was collected in three phases. Phase 1 entailed collecting formative data while following the Manual's linking process. Phase 2 involved post process reflections on the stages of the linking process by collecting summative data. Phase 3 comprises a reflection by members of the linking project on the validation approach implied in the Manual. The overall design is illustrated in the diagram in Figure 3.2 and discussed in detail below.

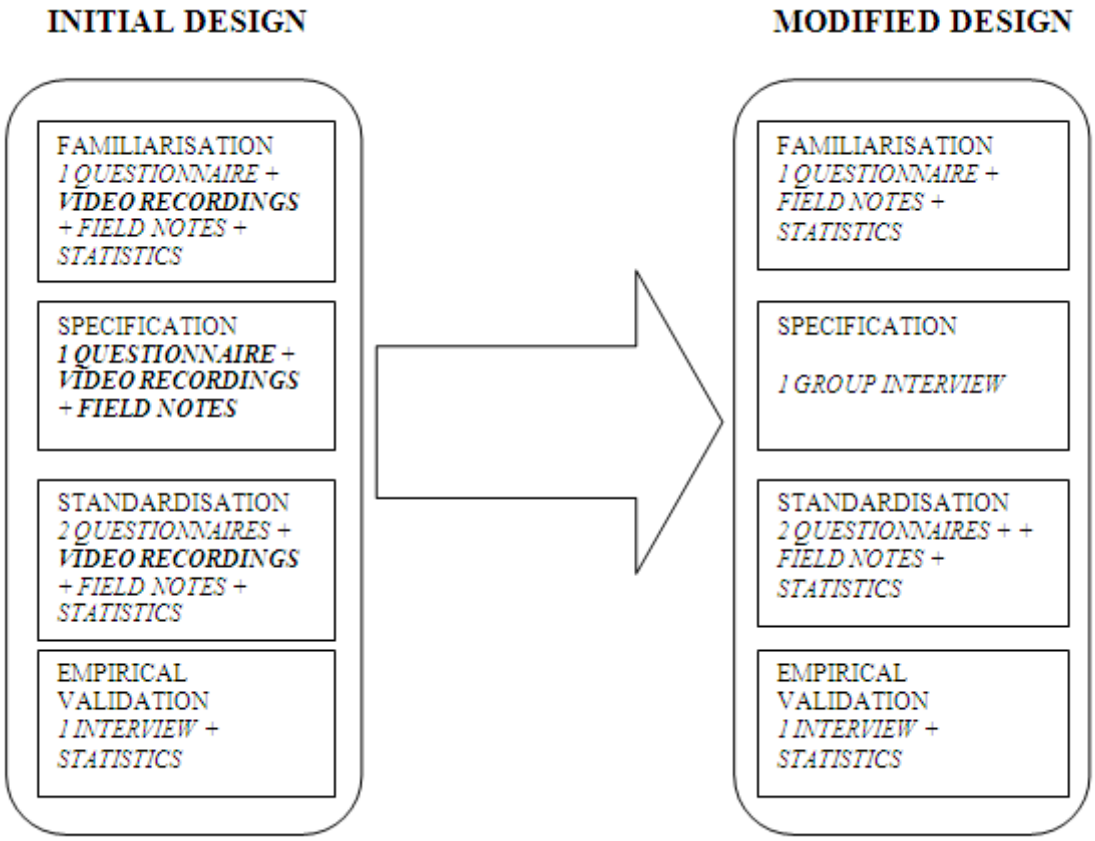
Figure 3.2 Data Collection Framework



3.3.1 PHASE 1 – Evaluation of the CEFR linking process

Phase 1 followed the four main stages of the CEFR linking project as suggested in the Manual viz. familiarisation, specification, standardisation, and empirical validation. Data at each of these four stages were collected differently according to the specific aims of each. Modifications were made to the initial design of this phase in the research, demonstrated in Figure 3.3, which are explained in the relevant sections that follow.

Figure 3.3 Initial and modified research designs – PHASE 1



3.3.1.1 Researching the familiarisation stage

The familiarisation stage, a requirement before the following two stages can be carried out (Council of Europe, 2003), ensures that participants gain an in-depth knowledge and firm understanding of the CEFR. The researcher proposes that understanding the CEFR and its levels through familiarisation means understanding the underlying language competences on which those undertaking CEFR linking studies are attempting to base their exam. Components of a test taker's language competence and strategic competence, as explained in Bachman's Communicative Language Ability model (1996), form aspects of context and cognitive validity as well as test taker considerations (Weir, 2005a). Even though the CEFR scales have frequently been criticized (e.g. Fulcher, 2004a, 2004b) for lacking theory, models of communicative

language competence (inter alia Bachman, 1990; Canale & Swain, 1980) influenced the initial stages of the development of the CEFR scales (North, 2000). In this regard, the descriptors include certain references to skills and underlying competences. Those involved in a linking study should have a firm understanding of the skills and competences encompassed in the CEFR descriptors, thus the levels. Familiarity with the CEFR will enhance the exam's consideration of aspects of a validation model in theory.

During this stage, the data collection aimed to determine the extent to which the CEFR had been internalised. The data provide confirmation or rejection of this. In addition, the Manual suggests familiarisation is ongoing throughout the linking process and, in this regard, it is a pre-requisite in the linking process. Institutions that work with people who are already familiar with the CEFR may choose to skip this stage. In this study, as most of the participants were not familiar with the CEFR, it was crucial to find out the extent to which the participants of the study grasped the scales they were asked to work with as a result of the activities carried out in the familiarisation stage. Lack of familiarity with the CEFR could have an impact on the results of the linking study. In other words, unless participants were familiar enough with the CEFR scales and levels to the extent that they could differentiate between the levels and apply the scales, the results of the linking study would have lost credibility.

To investigate levels of familiarity with the CEFR, and to triangulate data, four data collection methods were initially planned; questionnaires, video recordings, field notes, and quizzes.

a. Designing and analysing the questionnaire

A questionnaire was administered at the end of the first familiarisation session, it was deemed the best way of collecting data in a systematic and quick manner (Creswell, 2009) (See Appendix 3A for the familiarisation stage questionnaire). It reflected the issues that the researcher wanted people to reflect on viz.: prior knowledge of the CEFR; tasks assigned before the familiarisation stage; the effectiveness of the familiarisation session; the content material (CEFR); and the future needs of the participants. Table 3.1 presents an overview of the familiarisation questionnaire.

Table 3.1 Overview of the Familiarisation Stage Questionnaire

PART	FOCUS	NUMBER OF QUESTIONS
1	Familiarity prior to the project	6
2	Effectiveness of the pre-tasks	7
3	Effectiveness of the session and tasks used	9
4	Content material (CEFR chapters 4 and 5)	10
5	Future demands and needs	1 (open-ended)

In the questionnaire, a likert scale format was used to indicate degrees of agreement with the given statements, lending itself to quick and effective analyses (Anderson, 1998). Likert scale questionnaires allow for analysis of attitudes and opinions covering a negative-to-positive dimension, as well as comparability. In developing the questionnaire, as was the case with all the questionnaires used in the study, Anderson's (1998: 170) six essential steps, "determining the questions, drafting the questionnaire items, sequencing the items, designing the questionnaire, piloting and revising the questionnaire and developing a strategy for data collection and analysis," were followed.

The researcher first specified the objectives to be achieved through the questionnaire, followed by brainstorming of possible subheadings and statements for each sub-heading, and then designed the questionnaire. Muijs (2004: 51) suggests that "the single and most effective strategy to minimize problems is to make sure you pilot your instruments". A questionnaire needs to be piloted "using a group of respondents who are drawn from the possible sample but who will not receive the final refined version" (Cohen, et al., 2000: 261). However, the small number of project members (15 in total) who had no prior knowledge or experience of a linking or standard setting process meant that it was not feasible to pilot the questionnaire with the target population due to the nature of the project. In other words, without having gone through the stages of the project, the participants would not have had the background or the experience required to complete the questionnaire. Therefore, the questionnaire was given out for feedback to two colleagues who had previous experience of a linking study.

In the analysis of the questionnaire data, a descriptive data analysis using statistics, as described by Brown and Rogers (2002), was employed. In addition to percentage and frequency tables, basic statistics such as means, standard deviations and measures of skewness were also calculated so as to characterize the numbers in the data set in order to investigate the overall tendency of the group towards the areas pinpointed in the questionnaire.

Based on the feedback received at the end of the first familiarisation session that lasted over one and a half days, the familiarisation stage was extended and spread over an eight month period, with seven sessions in total. Extending the stage contributed to the participants' understanding of the CEFR, thus affecting the success of the proceeding

stages. Time intervals between sessions were arranged based on other work duties of the participants and institutional constraints such as exam and teaching times. In terms of content and tasks, the first session was the only one that completely reflected the suggestions in the Manual. As the aim was to assess the effectiveness of the suggestions made in the Manual, not those based on institutional needs requiring more time to get familiar with the CEFR, no further questionnaire was given in the later sessions. However, the extent to which the additional sessions helped the participants get familiar with the CEFR was investigated through the statistical analysis of quizzes administered as part of the familiarisation process during these extended sessions (See below). Two quizzes, one prior to the last familiarisation session and another before the writing standardisation, were administered for this purpose.

b. The use of video-recordings

Questionnaires have certain failings; ambiguities in questions may lead to misunderstandings that may not be detected in advance (Robson, 2002); and gaining a deeper understanding of questionnaire results and forming specific answers may be difficult (Muijs, 2004). In order to supplement questionnaires, sessions were initially video-recorded to keep a practical record of what was said, what happened, and what was presented in sessions; this was intended to permit the evaluation of participants, their responses in the tasks and discussions. Researchers cannot rely on memory to recollect conversations, as Sacks (in Silverman, 2000) cautions, only summaries of what different people said can be made. In this research, details of the discussions were crucial as they enabled the researcher to have follow-up interviews with specific people to clarify or further explore issues. Video recordings also display details of the setting and actions. In addition, they can be replayed for accurate transcription purposes.

However, video recordings were made for the familiarisation, specification and the first writing standardisation sessions but abandoned after the writing standardisation because some of the participants did not feel comfortable while being video recorded, as explored in Section 3.7.1.3, and the ones at hand were not analysed.

c. Gauging familiarity through field notes

While the video recordings were not used, field notes were kept, initially to supplement the video-recordings in case of electricity cuts, running out of tape, or sound quality (See Appendix 3B for an example of the filed notes coding scheme) even though adequate precautions were taken in advance to prevent such unpredictable events. Field notes later became important as video recordings were abandoned. At times, when the researcher was acting as a session leader, field notes were kept by the project leader.

Field notes consist of a researcher's detailed descriptive records of the research experience including discussions, reflections, descriptions and observations. Anderson (2002) regards field notes as important sources of data. In addition, Silverman (2000) points to a number of issues regarding field notes that contribute to their validity; the form in which the field notes are kept; what the researcher can see as well as s/he can hear; how the researcher is behaving and how s/he is being treated; and the expanding the field notes after the observation sessions.

Silverman's validity issues were addressed as follows. Before the project started, the project leader and the researcher got together to decide how to take field notes and came to an agreement that the notes would be kept in the form of running commentaries (1st issue). Who said what and when, including the nature of the task the participants were

asked to carry out (2nd issue), was noted down following the natural course of the discussions. Where possible, both the project leader and the researcher took field notes, using laptop computers to make the note-taking faster, thus reducing the possibility of missing out on information. With respect to the 3rd issue, as the researcher was a member of the project and also carried out all the tasks that the rest of the group had to do, she was not treated as an outsider making observations. As for the final issue, it was agreed that the person (the project leader or the researcher) who kept the notes, at times they both did, would expand them after the sessions.

Field notes were analysed using a coding system. Following Miles and Huberman's (1994) advice on coding, the research questions were used to create codes, which meant that the initial codes were created before the data analysis. For example, before the analysis of the field notes, the researcher considered the first research question "What part does linking an examination to external criteria have in the validation argument for that examination?" and the accompanying sub-questions and brainstormed words or phrases the interviewees could use under each validity type. For instance, participants might have talked about the CEFR descriptors and the scales, which are parameters of scoring validity. They might have also discussed the linguistic requirements of a task or in a descriptor, which are parameters of context validity and form a degree of 'cognitive load' (Shaw and Weir, 2009) that affects the cognitive processing demands imposed upon a test taker. Throughout data analysis, new codes emerged but there was no instance where a pre-determined code had to be discarded. The researcher devised codes that are semantically close to the terms they represent so that they could easily be used. The word 'scale' represented 'criteria' or the word 'structure' represented 'grammatical resources'. The list of codes prepared had definitions of each code as well. The next

stage was to prepare a coding system that consisted of the theme or the area the codes belonged to, the descriptions under that theme and the examples from the field notes (Appendix 3B for the coding scheme with relevant examples for the occurrences of the codes). As there was no other researcher involved in this study or anyone who could check the codes, the researcher undertook check-coding herself, which involved coding the same transcripts again to achieve a high percentage of code-recoding consistency, which is essential for reliability purposes. Different codes emerging the second time a set of data is coded means that the initial coding was unsuccessful and also raises questions regarding the second coding. The check-coding gave the researcher confidence that she was successful in her coding with an addition of only two new codes.

The next step was to perform pattern coding, which involves grouping the summaries gathered through coding into a smaller number of sets, themes or constructs. To do so, again the researcher kept to the sub-questions formulated for each research question. For example, the codes related to context validity were grouped together. Throughout the analysis the researcher looked for threads that would tie together sections of data and the research questions / sub-questions. Threads and links to the research questions were formed. Throughout coding, the memo writing strategy, which involves theorizing ideas for write up of codes and their relationships throughout the coding process (Miles & Huberman, 1994: 72), was also employed to help make sense of the data and remember initial interpretations at the final stage where conclusions are drawn.

d. Statistical measures used

Participant familiarity with the CEFR levels and descriptors was also investigated through five rank ordering tasks suggested by the Manual and two locally designed quizzes, one prior to the last familiarisation session and the second prior to the standardisation of writing. The results were used to decide whether participants as a group were ready to move on to the next stage. The rank ordering tasks determined whether the participants could follow the progression within the CEFR descriptors. The quizzes, added in the modified research design, determined whether the participants could recognise the features of each CEFR level. The quizzes also examined the usefulness of the additional sessions, designed in response to feedback, in enhancing familiarity. The data from rank ordering tasks and quizzes were statistically analysed (See accompanying CD Folder 1 Appendix 3C). Consistency and agreement among participants were analysed using Cronbach alpha and the intraclass correlation coefficient (ICC). Kaftandjieva (Council of Europe, 2004: 23) highlights in the Reference Supplement to the Manual that these types of analyses are inappropriate for standard setting purposes as "it is possible to have a perfect correlation of ± 1.00 between two judges with zero-agreement between them about the levels to which descriptors, items, examinees or their performances belong". The FACETS program which implements the many-facet Rasch measurement model, an extended Rasch model for dichotomous data to accommodate for the measurement needs of assessment situations (Linarce, 1989), was preferred. The advantage of the many-facet Rasch analysis is that all these facets can be compared on a common scale, which is called the 'logit' scale. Therefore, information can be obtained not only about examinee ability but also difficulty of tasks, severity and consistency of raters and the use of the rating scale (McNamara, 1996) (See Accompanying CD Folder 2 Appendix 3D for familiarisation

stage FACETS outputs). Used in similar studies (O'Sullivan, 2009a, 2009b, 2009c; Papagiorgiou, 2008), it allows for the performance analysis of the project members in terms of how well they know and use the CEFR scales and how consistent they are in their judgments, as evidence of the reliability of the judgment process. The statistical tools provide evidence as to the effectiveness of the suggestions made by the Manual in enhancing participant familiarity with the CEFR.

The CEFR levels were converted into quantitative data to facilitate statistical analysis through many-facet Rasch, and to calculate statistics such as ICC or Pearson Product Moment Correlation, which is common practice in CEFR linking studies (Papageorgiou, 2007a; 2007b; O'Sullivan, 2009a, 2009b, 2009c). The numbers assigned to CEFR levels are arbitrary and constructed for measurement purposes only; thus the number scale in Table 3.2 does not suggest equal distance between the levels. The levels included in the number scale might show variability depending on the purpose of the project and the stage of the linking process. For instance, at the familiarisation stage, all the CEFR levels were used whereas throughout standardisation, the number scale reflected only five levels; B1, B1+, B2, B2+ and C1. However, in most stages of the project the primary and plus levels ranging from A1 to C1 were used.

Table 3.2 CEFR Levels Converted Into Numbers

CEFR levels	Number Scale
A1	1
A1+	2
A2	3
A2+	4
B1	5
B1+	6
B2	7
B2+	8
C1	9

The statistical measures outlined in terms of their functions and how they were used at the familiarisation stage are presented in Table 3.3. The descriptive statistics were calculated in the session as they helped the participants see how the group perceived the CEFR descriptors and scales, also forming the basis for discussions to better understand the CEFR levels. The other measures viz. Cronbach Alpha, intraclass correlation, Pearson Product Moment correlation, and many-facet Rasch, were calculated after the session.

Table 3.3 An overview of the statistics used in the familiarisation stage

Statistics	Functions	As used in familiarisation
Descriptive statistics	The following were used in this research: • the mean i.e. the average score. • the mode i.e. the score obtained by the greatest number of people. • Minimum and maximum scores (CEFR levels in the case of this research).	Mean: In the familiarisation stage, the participants were asked to make judgments about a given descriptor. In this case, looking at the group mean helped the participants with the discussions and modifying their initial judgments. Mode: In making judgments about the level of a descriptor, the mode helped participants to see the group tendency which contributed to their decisions and the discussions. Minimum and maximum: These helped the participants to see the range of the judgments on a certain descriptor, which contributed to the discussions.
Cronbach Alpha	It provides a coefficient of each item (judge in this case) with the sum of all the other items.	It was used to provide evidence on and interpret participant reliability.
Intracalss Correlation Coefficient	It shows how the average rater agreed with all the others.	It was used to provide evidence on and interpret the agreement among participants.
Pearson Product Moment Correlation	It is used with interval and ratio data to show the linear relationship between two sets of data.	It was used to make judgments about the agreement among participants.
Many-facet Rasch	It helps looking at a score that is based on a number of facets (Linacre, 1994).	As well as identifying agreement among participants, it provides information about how consistent they were with their judgments.

3.3.1.2 Researching the specification stage

The specification stage ascertains whether an exam has been designed and produced following good practice. Exam coverage, what is measured, is reported in terms of the categories presented in Chapter 4 of the CEFR "Language Use and the Language Learner" and Chapter 5 "The user/learner's competences". The specification of exam coverage is considered to be a qualitative way of providing evidence of a link to the CEFR through "content-based arguments" (Council of Europe, 2003: 2). Chapter 4 – specification – of the Manual has a two-fold aim (ibid: 29). Firstly, it contributes to increasing the awareness among developers of quality language examinations of (1) the importance of good content analysis; (2) the use of the CEFR in planning and describing language examinations; and (3) the importance of relating language examinations to an international framework like the CEFR. Secondly, it defines minimum standards in terms of both the quality of content specifications in language examinations and the process of linking examinations to the CEFR.

The specification process is in two phases: a general description of the exam and a detailed description of the exam and requires filling in forms provided in the Manual. The general description involves undertaking a global analysis of the exam by filling in Forms A1 – A8 (Council of Europe, 2003: 34-41), which requires specification of the aim of the exam; domains involved; communicative activities tested; duration; test tasks; information provided for test takers and teachers; and measures used to report scores (ibid: 30-35). The detailed description of the exam entails filling in Forms A9 – A22 giving further details of each sub-test as regards Communicative Language Activities (CEF Chapter 4) and Aspects of Communicative Language Competence (CEF Chapter 5). The results are presented graphically demonstrating the exam

coverage in relation to CEFR levels (ibid: 30). The list of forms which were completed for this case study relevant to the COPE Exam in general and the Reading and Writing papers in particular are presented in Table 3.4. The end-product of the specifications stage is a specification in the form of a report that makes examinations more transparent to users of exam results and test takers (ibid: 29).

Table 3.4 Overview of the forms completed at the specification stage for reading and writing

FORM	FOCUS
A1	The general exam description
A2	Test development
A3	Marking
A4	Grading
A5	Reporting results
A6	Data analysis
A7	Rationale
A8	Impression of overall examination level
A10	Reading comprehension
A14	Written production
A19	Aspects of language competence in reception
A21	Aspects of language competence in production
A23	Graphic profile of the relationship of the examination to the CEF levels

The Manual suggests that the team responsible for the examination complete the specification forms (ibid: 29). In this case, the whole project group participated in this process. The project leader and the researcher perceived this as crucial to achieve an outcome that reflected the ideas of all the project members. The members, led by the project leader, first came to a common understanding regarding the terminology in the forms through discussion of every question in the forms and then filled in the written production (A14) and reading comprehension (A10) forms. These were completed during a session, whereas the language competence forms were filled in outside a session, as the institution could not bring the project members together once again prior

to the standardisation stage due to the pressures of their regular duties. The time of the standardisation stage had been set as external experts had been invited to take part in this stage. Thus, the project members could not hold another session for the specification stage.

It had been planned to research the specification stage in the same way using similar tools viz. questionnaires, field notes and video-recordings to gather data to inquire into the contributions of the specification stage to the validity of the COPE examination and its level. One questionnaire was to be administered mainly to address the third research question focusing on the institutional implications. Field notes and video-recordings would be used to seek answers to the second research question involving the validity of the reading and writing papers, and the third research question. However, data collection proved problematic for logistical and people-related reasons explained in what follows; therefore, these tools could not be used.

Initially, half a day was allowed for the specification session but it became apparent that more time had to be allocated as clarifying the terminology in the specification forms took longer than expected and the participants needed sufficient time to discuss the competences measured through the COPE papers. However, the institution could not accommodate a further session due to the needs of daily operations. For the session held, although video-recording was done and field notes were kept, they did not reflect the whole specification stage of the linking process as project members were not able to satisfactorily complete the process. For the same reason, the questionnaire could not be distributed at the end of the specification session as initially planned.

Because the group could not meet again to complete the specification forms, the project leader filled in the language competence forms with the researcher and sent them to the participants for confirmation. This had to be done several times until the forms were completed. During the process of checking and completing the language competence forms by the project members, a number of issues arose; firstly, a lack of agreement arose on what some of the questions in the forms meant; secondly, some project members did not return their forms; and, thirdly, others pointed out that they were not clear about how to complete the forms. For example, some had the target test takers, that is their own students, in mind while filling in the forms whilst others could not easily relate the exam to the reference CEFR tables given in the forms. From the large group of 12 project members who were involved in the specification stage, the most useful data (forms completed in detail with relevant and expanded justifications showing thorough analysis of the exam) came from the three members of the team who were heavily involved in testing, who had experience of writing test specifications, and who were particularly familiar with the COPE exam.

a. Specification stage interview

Since no other formal session took place, further recordings or field notes were not made. Furthermore, analysing the data from a questionnaire to three people would have posed serious threats to the credibility of the data. It was decided, therefore, to conduct a group interview in the hope of producing a large number of responses, generated through discussions, in a short time. (Cohen, et al., 2000). The group interview was conducted with the three people who were deemed to have contributed most fully to the specification stage, as mentioned in the previous section, and as such were felt to be best positioned to participate in the interviews. The format was devised to cover all the

research questions of this study through the interview (See 3E for specification stage interview coding scheme). The discussions were particularly important in that the interviewees exchanged ideas and discussed one another's opinions.

The validity and reliability of the specification stage interview were addressed through a set of principles. Cohen et al. (2000: 121) advocate that the most practical way of achieving greater validity is to minimize the amount of bias resulting from the following issues:

(a) the attitudes, opinions, and expectations of the interviewer;

(b) a tendency for the interviewer to see the respondent in her own image;

(c) a tendency for the interviewer to seek answers that support her preconceived notions;

(d) misperceptions on the part of the interviewer of what the respondent is saying;

(e) misunderstandings on the part of the respondent of what is being asked.

In order to raise her consciousness about the issues involved in carrying out interviews and to be neutral and objective, the researcher had discussions on how to conduct interviews with her supervisors and used the input received in the research methods course she took as part of her programme (a and b). In the interview, the researcher had some pre-determined questions in hand and used those to guide the interview but allowed the interviewees to lead the discussions so as to prevent herself from attempting to elicit preconceived information (c). Ensuring that what the interviewee said was perceived accurately by the researcher entailed summarizing the discussion and sending or showing it in writing to the interviewees for confirmation and allowing them to challenge what she had written (d). As for misunderstandings on the part of the

interviewee, semi-guided interviews made it possible to intervene and make clarifications when a concern over misunderstanding of a question arose (e). As for reliability, a highly structured interview where the same sequence of words and questions are followed for each interviewee is a way of controlling reliability of interviews (Oppenheim, 1992; Silverman, 1993), which also guarantees that every interviewee is presented the same questions in the same way. However, a semi-structured interview was used because it allowed the interviewees to elaborate on issues and would not limit the depth and breadth of the responses (Dörnyei, 2007). This also increases the validity of the data collected.

In terms of data analysis, the first step was to transcribe the interviews soon after they had been conducted. Secondly, to aid future analysis, coding was required and the same steps, described under 3.7.1.1 for the analysis of field notes, were followed. The codes generated for the field notes were used for the analysis of the interview (See Appendix 3E for the coding scheme with relevant examples for the occurrences of the codes). The researcher undertook check-coding herself, with no additional codes added, increasing confidence in the validity of the codes. Pattern coding, grouping similar points into themes, and memoing, theorizing write-up ideas, were also carried out for the specification stage interview.

b. Specification forms and participant agreements as data

Eight specification forms were completed: three by the whole group and the remaining five by five people, including the project leader, researcher and the three people mentioned above. The completed specification forms were later used as data for validation purposes (See accompanying CD Folder 3 Appendix 3F for the completed

forms). They served as evidence of the validity of the exam itself and were used to corroborate the exam level established at the standardisation stage. Because only three forms were completed as a group effort, the researcher changed the way she researched the specification stage. The researcher initially intended to use field notes, video-recordings and a questionnaire at the specification stage but then decided to carry out an interview with the three people who contributed most to the completion of all specification forms. In addition, statistical data from the session carried out were available in the form of percentages reflecting the agreement among the project members on the levels assigned to the competences measured through the COPE examination.

3.3.1.3 Researching the standardisation stage

The aim of the standardisation stage, as the name suggests, is to ensure that the CEFR levels are implemented consistently by the participants involved in the linking process (ibid: 65). The standardisation process comprises four steps:

- Familiarisation: So as to be thoroughly familiarized with the CEFR levels, the participants are asked to do similar activities to those in the familiarisation stage as part of the standardisation stage.
- Training: The participants use the standardized exemplars provided by the Council of Europe to assess learner performance for productive skills or assess the difficulty of items for receptive skills in relation to CEF levels. They need to justify their judgments, and thus acquire experience in relating performances or items to CEF levels. Reaching consensus as a group is crucial for training purposes (ibid: 70-71).

- Benchmarking performances: This is the application of the consensus reached at the training to the assessment of local samples and involves activities similar to the ones carried out in the training process. The outcome is a set of locally standardized items or performances.
- Standard setting: The process of setting cut scores for the different sub-tests in an exam in relation to CEF levels. The initial judgments need to be confirmed with empirical evidence gathered from live test administration. This then leads into investigation of internal and external validity, dealt with at the empirical validation stage (ibid: 71).

a. The use of questionnaires

Questionnaires (one for reading and one for writing) were designed to evaluate the effectiveness of the standardisation sessions (See Appendix 3G for the standardisation stage reading and writing questionnaires). Participants were asked to indicate their opinions on whether the standardisation stage had an impact on the validity and reliability of the exam in general terms without going into different aspects of validity. They were also asked to evaluate the usefulness of the exemplars provided by the Council of Europe to facilitate standard setting. Furthermore, the questionnaire was used to evaluate the validity of the standard setting itself. Evaluation of the standardisation itself was important because, unless the sessions had been perceived as being effective, then participants would not have been in a position to make meaningful comments in return about the impact of this stage on the exam and the institution. If a standard setting session lacks quality and fails to achieve its aims, the participants involved may not have a firm understanding of what standard setting involves. Thus, any conclusions regarding the validity of the examination as a whole might not be valid.

In addition, preliminary answers to the research questions were sought, as it was crucial for them to go through the whole linking process to have a better idea of what CEFR linking meant and what it involved.

b. Video-recordings and field notes

Video-recordings, initially used, had to be abandoned after the first writing standardisation based on informal feedback from a number of project members who objected to being recorded. They reported feeling uncomfortable as it forced them to be more considering of their language, causing them to be less spontaneous and more contrived in voicing their opinions. A further concern lay in the fact that participants were worried that their managers might gain access to the tapes and could monitor and judge the behaviour or attitudes they expressed in the sessions. Thus the researcher abandoned the recordings for ethical reasons. Furthermore, the danger that the data generated from video-recording did not truly reflect participants' opinions and would yield distorted data and pose a threat to reliability of the findings, the ones made during the first stage were not used either. As a result no video-recordings were used in the study. The fact that video-recordings were no longer continued in return increased the importance of field notes. The field notes at this stage were kept as records of what was done, in the same fashion as for the familiarisation stage (See Appendix 3H for the standardisation stage field notes coding scheme), and analysed in the way previously described for the analysis of interviews in section 3.7.1.2.

c. Statistical measures

Setting reliable cuts cores is crucial in the validation of an exam. A cut score is a selected point on the score scale of a test that determines whether a particular test score

is sufficient for some purpose (Zeiky & Perie, 2006) or for a performance standard. In the COPE linking project, the performance standard was CEFR B2 level and the aim of the standard setting was to identify the score on the COPE – the cut score – that corresponds to the B2 level. Different standard setting methods such as the Angoff method (1971) and the Examinee-Paper Selection method (Hambleton, et al., 2000) were used to arrive at the cut score. Scoring validity needs to be investigated to determine the reliability of the cut score established. Scoring validity in standard setting involves analyses of judge severity and consistency as well as the use of criteria. In order to explore whether the cut scores (for reading and writing) set at this stage were trustworthy and that the scoring was carried out to acceptable standards, the many-facet Rasch model was employed, to analyse the judgements of the participants on the items or written samples, for a number of reasons mentioned above in 3.7.1.1. First of all, it allows analysis of observations resulting from more than two facets, ie. difficulty and discrimination. Secondly, it places all facets on a common scale. Therefore, information can be obtained not only about items or sample performances used for standard setting but also the severity and consistency of the judges and the use of the CEFR scales, which might have influenced the benchmarking, that is, the proposed cut scores, thus the validity of the COPE examination (See accompanying CD Folder 4 Appendix 3I for standardisation stage FACETS outputs). The statistics used at the standardisation stage are the same as the ones used in the familiarisation stage.

3.3.1.4 Researching the empirical validation stage

Empirical validation provides evidence that the exam is valid and that claims made at the end of the specifications stage and the standardisation stage of the linking process are reliable and can be confirmed by reference to other criteria, such as another test

linked to the CEFR, through the use of statistical tools (Council of Europe, 2003: 2). Empirical validation is seen in the Manual to comprise of two parts: internal validation and external validation, a limited and outdated approach to validation (O'Sullivan, 2009a). In this research, as prescribed in Weir's validation framework, the internal validation model in the Manual is considered to be a parameter of scoring validity and the external validation is regarded as a parameter of criterion-related validity.

Internal validation, as seen in the Manual, is about establishing the quality of a test, a pre-requisite for linking to the CEFR. The internal validation procedures described in the Manual are based on two main classes of statistical test theories. The Manual suggests that a number of empirical analyses, namely, Classical Test Theory, Item Response Theory, qualitative, generalisability and factor analysis be carried out where relevant to provide evidence for internal validation.

External validation, on the other hand, is essential in verifying the relationship of the cut scores set for an exam with the CEFR levels themselves through empirical evidence (ibid: 108). It mainly involves correlation; that is, correlating the scores on the exam in question, COPE in this case, with those of a measure of the intended construct, and matching classifications based on the exam under study and classifications made by the external criterion. This entails converting a quantitative test score into a qualitative category represented by the CEFR levels (ibid: 109), for example, which CEFR level does a score of 30 out of 50 in a given test correspond to?

At the empirical validation stage, the researcher aspired to find out whether the statistical analyses carried out to validate the COPE reading and writing papers and the

linking process had any impact on the exam with respect to scoring and criterion-related validity. This stage had to be analysed in a different way to the other stages of the linking process. In terms of internal validation, it does not require the involvement of the project participants as it relies solely on the statistical analysis of an exam in order to collect evidence on the validity of the exam. In terms of external validation, teacher judgments were a part of the process, and, for this purpose a different group of teachers than the ones taking part in the linking project were trained in the CEFR. It was important that these teachers taught full time so that they could monitor and make judgments about their students' language proficiency levels prior to the COPE examination. Once the COPE was administered, correlation analyses of COPE reading and writing papers with teacher judgments were carried out. In addition, a comparison between teacher judgments and student COPE scores was made. This information was shared only with the project leader for reasons of confidentiality. As the analyses were carried out on an experimental basis, the cut score established as a result of the analyses could not be shared with others before it was confirmed. In other words, information regarding the analyses results was not shared with others until the level of the exam in relation to the CEFR was set solid. (See accompanying CD Folder 5 Appendix 3J for the empirical validation stage FACETS and QUEST outputs).

For research purposes, besides the use of statistical data in investigating whether the statistical methods suggested in the Manual contribute to the validity of the COPE examination, an interview with the project leader was scheduled after the empirical validation stage to find answers to the research questions regarding this stage (See Appendix 3K for the empirical validation stage interview coding scheme). The interview was semi-structured, in that questions were predetermined (Robson, 2002). A semi-structured interview was chosen because the project leader was not an expert in

statistical test analysis, thus would need help throughout the interview in terms of clarifying the interview questions or the terminology used in the questions if deemed necessary. Semi-structured interviews allow for explanations during the interview, which enables obtaining the required information by avoiding misinterpretations, in this case the terms in the interview questions could be clarified if necessary. The interview data were analysed in the same way as the interview at the specification stage was analysed.

3.3.1.5 Summary of the issues arising from Phase 1 research design

The initial design had to be substantially modified (Figure 3.2), as explained above, due to a number of issues. Initially, three data collection instruments (questionnaires, video-recordings and field notes) were to be used at all stages of the linking process except for empirical validation. However, video-recordings had to be taken out since some participants expressed discomfort in expressing their views in front of a camera, as discussed above. Field notes are available for familiarisation and standardisation sessions but only for a small part of the specification stage due to the problems mentioned specifically in section 3.7.1.2, and some participants not being very clear on how to fill in the forms. Therefore, instead of a questionnaire and field notes, a group interview was carried out with three people for this stage. In addition to questionnaires and field notes, statistics also provided data. Data for the empirical validation comes from statistical analysis and an interview with the project leader.

3.3.2 PHASE 2 – In-depth analysis of the CEFR linking process

This phase of the research acted as a means to gain a deeper understanding of the CEFR linking process through gathering further quantitative evidence post the treatment, i.e.

after the project members had gone through all the stages of the CEFR linking process. Phase 2 aimed to collect evidence on all aspects of validity (Research Question 1), with a particular focus on reading and writing skills (Research Question 2) and to what extent, if at all, they were considered throughout the linking process. It also aspired to gather evidence regarding the contributions of the linking process to pinpointing areas in the COPE examination in order to adjust it to fully reflect the intended language proficiency level.

A questionnaire was preferred at this phase as it enables the use of standardized questions, which allow for comparability (Muijs, 2004; Cohen et al., 2007) and allow for a large amount of information about all parameters of validity to be collected in a systematic and quick way (Creswell, 2009: 146). Aspects of validity presented by Weir in his validation frameworks (2005a) formed the source that informed most of the questions. The questionnaire, in a multiple choice format, comprised seven parts (See Appendix 3L for Phase 2 questionnaire). Parts 1 to 6 focused on aspects of validity, e.g. context validity, scoring validity. The number of questions in each part reflected the number of parameters in Weir's validation frameworks (2005a) for that specific aspect of validity. Each of the parameters under an aspect of validity was turned into a question. Questions were prepared for reading and writing separately where they had different parameters; when they had shared parameters, the same questions were used but the project members were asked to answer them for both reading and writing independently. The project members' task was to indicate whether or not and at what stage of the CEFR linking process each parameter of aspects of validity was considered. Table 3.5 shows an overview of the questionnaire.

Table 3.5 Overview of Phase 2 Questionnaire

PART	FOCUS	TOTAL NUMBER OF QUESTIONS	
		Reading	Writing
1	Test taker	3	3
2	Context validity	19	19
3	Cognitive validity	14	15
4	Scoring validity	7	10
5	Consequential validity	3	3
6	Criterion-related validity	4	4
7	Institutional implications	5	5

Whereas Parts 1 to 6 targeted research questions related to validation, Part 7 was specifically incorporated to address the last research question regarding institutional implications, with five questions that aimed at finding out whether the CEFR linking process contributed to the understanding of the level of the COPE reading and writing papers and whether this level was suitable for academic study.

The questionnaire was administered about two weeks after the linking process ended and required respondents to think back at the stages of the linking process and identify aspects of validity considered at each of the stages. As the project dated back two years, details of the stages such as the session notes, tasks carried out and the slides used were made available to the participants to aid them; however, no one needed them. A few participants wanted to make sure they knew the names of the stages, i.e. familiarisation, specification, etc. correctly as the names were key to the completion of the questionnaire.

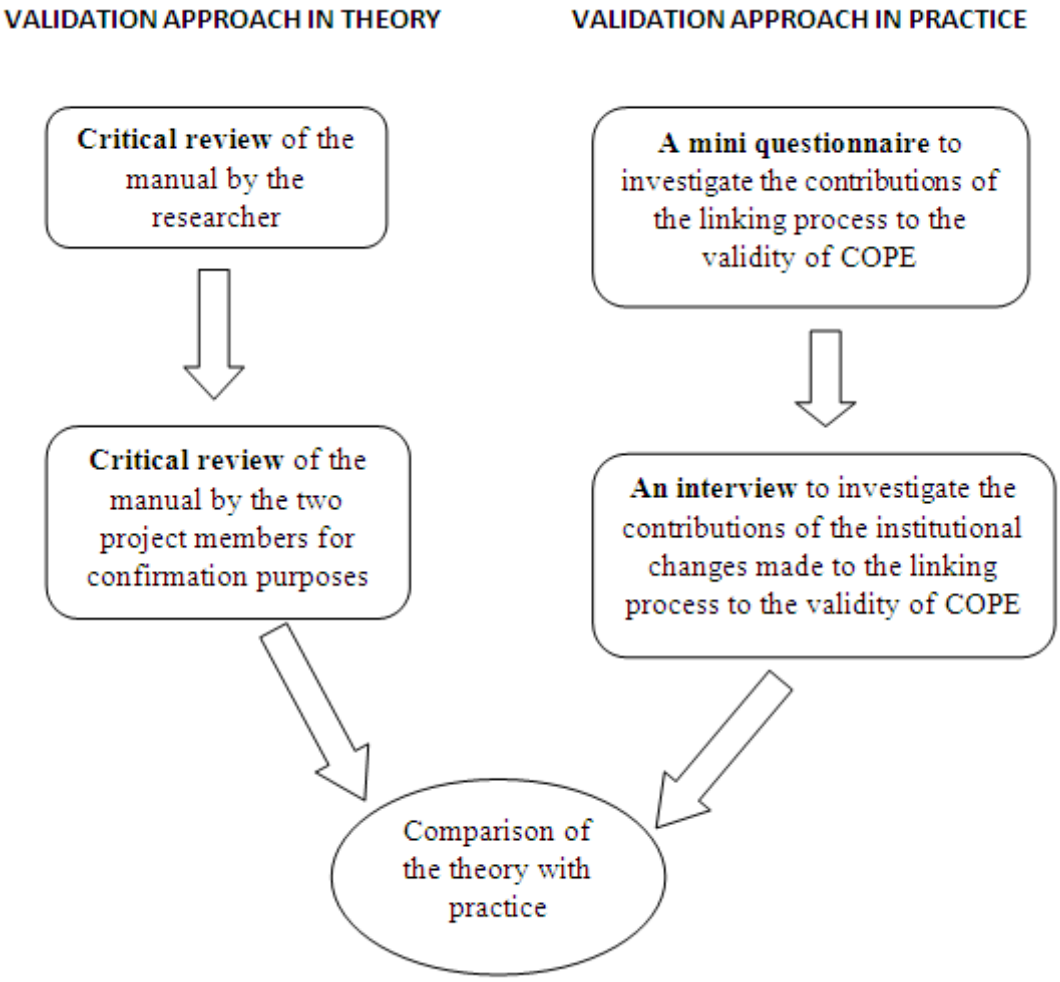
In the analysis of the questionnaire data, a similar analysis to that carried out in Phase 1 of this research was employed, i.e. descriptive data analysis using statistics of percentages, frequency tables, means, standard deviations and measures of skewness to characterize the numbers in the data set.

3.3.3 PHASE 3 – A review of the Manual**'s approach to validation**

In Phases 1 and 2 data were collected on the application of the theory put forward in the Manual, incorporating some institutional modifications such as an extended familiarisation stage. Phase 3 of the research had two purposes; firstly, to critically review the Manual as a theoretical model for linking with the aim of identifying to what extent the model put forward by the Manual catered for aspects of validity as defined by Weir (2005a) (reasons for the preference of Weir's framework over others were explored in Chapter 2). Secondly, it aimed to look at the validation of the COPE from the perspective of what the Manual suggests, which involved evaluating whether the suggestions in the Manual contributed to the validity of the COPE examination, as well as the contributions of the institutional adaptations made to the suggestions in the Manual. Figure 3.4 presents an overview of the research methodology used in Phase 3.

Throughout the project, additions or changes were made to the procedures suggested in the Manual in response to contextual needs, institutional restrictions and an increase in the institution's knowledge of the CEFR and linking. These changes might have affected the efficiency of the procedures suggested in the Manual; therefore a need arose to analyse the methodology of linking in order to investigate how the changes or additions made in the BUSEL context influenced, or not, the success of the linking process. With this in mind, the researcher carried out a critical review of the Manual using a chart (See accompanying CD Folder 6 Appendix 3M), described in what follows, then gave a mini questionnaire (Appendix 3N) to three project members who worked in the Testing Unit and thus had access to the COPE examination and confidential data related to the examination, and had individual interviews with the three members to follow-up on the results of the questionnaire.

Figure 3.4 Overview of the research methodology in Phase 3



The chart was designed to review the approach to validation in the Manual and different versions were produced for reading and writing each consisting of the following:

- Column 1 – aspect of validity (as presented in Weir);
- Column 2 – parameters of each aspect of validity (as presented in Weir);
- Column 3 – how this parameter/aspect of validity is tackled in the Manual;
- Column 4 – a comments section for people who checked the document, for confirmation purposes.

The researcher reviewed the Manual chapter by chapter looking for parameters of each aspect of validity. Sometimes the evidence was explicit in the Manual such as the type of analysis that needs to be carried out for internal validation. Sometimes, the researcher had to look for clues similar to the codes she used in the analysis of the interviews or the field notes. For instance, with respect to criterion-related validity, the evidence came from the specification forms where users were encouraged to question the competences measured in a test. It should be emphasised that Phase 3 was a review of the approach to validation as stipulated by the Manual, not a review of the Manual itself.

In order to address issues of quality, the resulting chart of the critical review was given to two people, who were asked to add, delete or change anything in the resulting chart, as well as confirm what the researcher had found. The two criteria used in the selection of these people were (1) to be a project member, which meant having taken part in the research, and (2) to be familiar with Weir's framework, used as the basis of analysis. These two people were different from the ones who took part in the interviews in Phase 2.

The mini questionnaire, given after the critical review was completed, aimed to identify what type of validity evidence was gathered for the COPE examination as a result of the CEFR linking project. It consisted of nine questions related to validity and for each question, the respondents were asked whether the CEFR linking process helped collect evidence towards a given aspect of validity.

- Column 1 – aspects of validity and parameters
- Column 2 – whether an aspect was realised or considered as a result of the linking of the reading paper

- Column 3 – whether an aspect was realised or considered as a result of the linking of the writing paper

The respondents were also asked to use the back of the questionnaire if they had any comments.

Peer feedback on the questionnaire was collected prior to administration and the results were analysed by totalling up the number of respondents under each column.

The interviews, carried out after the completion of the mini questionnaire, aimed to further investigate the results of the questionnaire and the extent to which the modifications made to the suggestions of the Manual contributed to the success of the COPE CEFR linking and the validity of the examination. A chart with the stages of the linking process and the modifications made to the suggestions by the Manual was used as the basis of the interviews and the participants were asked to comment on the areas covered in the chart in terms of the contributions of the modifications to the validity of the COPE examination. The analysis of the interviews was done in the same way as the analysis of the interviews in Phase 1. Codes were generated from the research questions and check-coding, showing the researcher was successful, was carried out.

Chapter 6 of this research aims to report the findings of this phase. In that chapter, comparisons between theory and practice will also be made based on the data collected from the critical review of the Manual and the interviews, to shed light on the extent of how the adjustments made to the Manual approach throughout the study have affected the running and the results of the project in terms of reading and writing from a validation perspective.

3.3.4 Summary matrix

The summary matrix below (Table 3.6) presents an overview of the tools used in this research. At different stages of the study, different tools were used to gather data. The links between the tools and the research questions they were used for can be seen in this matrix.

Table 3.6 Summary Matrix Linking Instruments in Each Phase to Research

Phase & RQs / Instruments	PHASE 1			PHASE 2			PHASE 3		
	RQ1	RQ2	RQ3	RQ1	RQ2	RQ3	RQ1	RQ2	RQ3
Questionnaires	√	√		√	√	√	√	√	√
Field notes	√	√							
Interviews	√	√	√						
Statistical Data	√	√	√	√	√	√			
Critical Review							√	√	

3.4 Ensuring ethical integrity

Ethics "concerns the system of moral principles by which individuals can judge their actions as right or wrong, good or bad" (Denscombe, 2002: 175). The individual lies at the heart of this definition, which therefore implies that the feelings, rights and welfare of the participants ought to be taken into consideration by the researcher. From a moral point of view, Denscombe (ibid) also suggests that researchers are obliged to stay within the law and cultural norms of the society within which they conduct the research.

Fundamental concerns for ethical research according to Borg (2006) are as follows:

- obtaining informed consent
- avoiding harm
- avoiding deception
- maintaining confidentiality

- maintaining anonymity

Ethics in research entails being clear about the nature of the agreement you have entered into with your research participants. The agreement should encompass the above points and, at the design stage of a proposal, a researcher is required to consider the implications of the topic and the methodology. Every researcher's ultimate aim is to put together a well-designed research methodology, however, what they ought to consider while doing so is to bear in mind the ethical side of research, that is, the welfare of the participants.

Informed consent

Ethical research involves getting the informed consent of those you are going to interview, question, observe or take materials from. It also involves reaching agreements about the uses of this data, how its analysis will be reported and disseminated, and keeping to such agreements when they have been reached (Bell, 2002: 39). Anderson (2002) considers informed consent as the most fundamental principle for ethical research and indicates that the participants involved in the research must be informed of the nature and purpose of the research, its risks and benefits, and must consent to participate without being forced. The researcher used the informed consent form designed by Roehampton University for the participants (See Appendix 3O for a copy of the consent form). The form briefly outlined the purpose of the research and its procedures. The researcher distributed the consent form to the participants in the very first meeting of the project. The initial consent form could not include all the details of the research process as the project was foreseen to last more than two years. Changes or additions to the research procedures were inevitable given

the nature of the project. Whenever a procedure such as interviews, questionnaires or video-recordings was introduced, the researcher informed the participants and asked for their consent in writing so that they could have the option of dropping out at any time.

Avoiding harm

Descombe (2002) points out that researchers need to be sensitive and think ahead to avoid any aspects of the research that could cause distress or physical discomfort to the participants. This study only required participants to do questionnaires or sit in an interview regarding topics surrounding the field of language assessment and testing, and the need to avoid harm became an issue once throughout the study when some participants felt uncomfortable being video recorded. Therefore, the researcher decided not to use video recordings (see section 3.7.1.3), which is an example of not wishing to use an approach that participants might conceive as potentially harmful. It may be that we need to broaden our description of ethical research behaviour to include the notion of perceived harm as well as actual harm to individuals. If an individual thinks they might be harmed, it would not be ethical to continue with the procedure simply because it does not involve actual harm.

Avoiding deception

Avoiding deception in research involves researchers' being honest and open about who they are and what they are doing, and they should not rely on deception as a means to get the information they are seeking (Descombe, 2002). In the case of this study, as the researcher was an instructor in the school and thus a colleague of the project members, her identity was not a cause of concern. Furthermore, she clearly wrote down or shared

verbally the aims of the research methods every time she collected data through the use of questionnaires, interviews, etc.

Maintaining confidentiality and anonymity

Maintaining confidentiality and anonymity was treated with caution throughout the study. These two aspects of ethical research are seen as interlinked by Anderson (2002), as according to him, confidential information does not only mean that the identity of the participants will remain anonymous but it also suggests that the reader of the research will not be able to deduce the identity of the participants. In order to maintain confidentiality and anonymity, the questionnaire and interview results were reported in an anonymous format and where quotations from interviews or video-recordings were used, the participants were given codes, which could only be traced back by the researcher.

3.5 Summary

In this chapter, the argument for a case study with a mixed methods approach was made (Section 3.2). Following on from this, the research design with respect to the instruments used, the rationale for their use and quality issue for each instrument were discussed (Section 3.3) Finally, issues regarding research ethics were discussed (Section 3.4).

The next chapter will discuss the first of the three phases of this study; the formative evaluation of the CEFR linking process.

CHAPTER 4

PHASE 1 – EVALUATION OF THE CEFR LINKING PROCESS

4.1 Introduction

This chapter aims to present Phase 1 of the research as outlined in section 3.7.1 of the previous chapter. It follows the order of the stages of the linking process as suggested by the CEFR Manual; namely familiarisation (in section 4.2), specification (in section 4.3), standardisation (in section 4.4) and empirical validation (in section 4.5) to enable the evaluation of each stage separately. For each of these stages, the suggested approach to linking examinations to the CEFR as described in the Manual is presented, followed by how each stage was undertaken in the COPE CEFR linking project, and how the stages were researched for this case study. Finally, the data are analysed for each stage and findings are presented. A critical review of the Manual approach to linking is given in the summary sections of each stage and in the last section of the chapter (section 4.6), overall conclusions drawn from Phase 1 of the research are presented.

4.2 Familiarisation stage

4.2.1 The Manual approach

The familiarisation stage aims to ensure that participants in the linking process gain an in-depth knowledge of the CEFR, an essential requirement before carrying out the subsequent two stages, viz. specification and standardisation (Council of Europe, 2003). The manual suggests two groups of familiarisation activities; namely 'Introductory activities' and 'Qualitative analysis of the CEFR scales'.

The former activities require participants to read a number of questions and reflect on them. The questions involve considering issues such as the purpose, approach to teaching, and expectations of the learners in question. They also require discussion of the CEFR global scale, which summarizes the 6 main levels for all skills. In addition, a discussion of the ELP self-assessment grid, a core element of the Swiss model of the European Language Portfolio (Lenz & Schneider, 2002: 69-70), is also recommended. Although not stated categorically, these activities aim to bring the key features of the CEFR levels in terms of progression to the attention of the participants.

The latter activities, called "Qualitative Analysis", require participants to sort CEFR descriptors into piles by level and rank order them. The Manual also recommends reconstructing CEFR Table 2 (the Swiss ELP) and sorting the DIALANG descriptors for each skill into levels (Council of Europe, 2003: 238-243). DIALANG is an online assessment system in 14 European languages intended for language learners who want to obtain information about their language proficiency. The DIALANG descriptors are one of the sets of descriptors that complement the CEFR itself (Council of Europe, 2001). The essence of these rank ordering and reconstructing activities, although not stated specifically, is to further strengthen familiarity with the CEFR levels and scales through working with the descriptors rather than discussing them.

The Manual advises users to choose at least one activity from each group of activities at the familiarisation stage and recommends revisiting these activities before the specification and standardisation stages. The empirical validation stage is not included in the familiarisation process, as it does not require people involvement unless teacher

judgments are used to validate the recommended cut scores. It requires a number of statistical analyses of the live test data.

4.2.2 Overview of the familiarisation stage of the project

As mentioned above, by including a Familiarisation stage, the Manual is making a statement that the group of participants in a CEFR linking project should have an in-depth knowledge and understanding of the CEFR (Council of Europe, 2003). As borne out in the familiarisation questionnaire at the beginning of the project, only five of the project members were somewhat familiar with the CEFR with a group mean of 2,4 on a scale of 5 (See section 4.2.4.1 for details of the questionnaire findings). Therefore, the project members and members of the senior management agreed to extend the period of time allocated to this most important aspect of the linking process during the introduction to the CEFR and the linking.

The Manual suggests the timetable for the familiarisation stage as being approximately three hours (ibid). However, two decisions were taken regarding the time frame of the familiarisation stage; the first one, prior to the familiarisation session. Considering the inexperienced profile of the group, the project team decided on 1.5 days of familiarisation to ensure a deeper and firmer understanding of the CEFR in general and help project members relate CEFR levels and descriptors to their own teaching experience with reference to BUSEL levels, thus going beyond a simple familiarity with the Illustrative scales (Council of Europe, 2001).

The second decision was taken to extend the familiarisation stage even further with the project team based on concerns that came to light throughout the introductory workshop

and the feedback received through a questionnaire. The participants' responses, presented fully in section 4.2.4.1, showed that they had difficulty understanding the rationale behind the CEFR. They were not convinced that the CEFR could be used in any context, and thus relating it to their own context was difficult. Two people also indicated that they wanted to question the descriptors in more detail before moving on to the other stages. (See Appendix 4A for the results of the questionnaire). The results of the questionnaire also showed that the time framework in the Manual was underestimated and institutional constraints such as having to work with local people, who are not familiar with the CEFR were not taken into account. In cases where people in an institution are not familiar with the CEFR, the researcher suggests that outside experts can be invited to take part in the project, facilitating the familiarisation stage.

The familiarisation stage of the BUSEL linking project, extended over eight months with a total of seven formal sessions, is explained in detail below.

Session 1 – Familiarisation session following the Manual format (app. 10 hours)

Session 2 – Update meeting (app. 2 hours)

Session 3 – Berlin update and article discussion (app.2 hours)

Session 4 – Standard setting simulation (app. 3 hours)

Session 5 – Overview of the study on the construction of the CEFR descriptors (2 hours)

Session 6 – In-depth analysis of CEFR descriptors (3 hours)

Session 7 – Meeting to discuss issues regarding embedding the CEFR into the BUSEL context (Invited speaker Prof. H.A.L. John DeJong) (2 hours)

4.2.2.1 Session 1 – Familiarisation workshop following the Manual format

The first session, in the form of a workshop delivered over 1.5 days, followed the activities suggested in the Manual viz. Introductory Activity (b) (Council of Europe, 2003: 26) and Qualitative analysis of CEF scales (d) and (e) (ibid: 27). (See Appendix 4B for the outline of the workshop and session notes). The participants were given a pre-session task that asked them to read CEFR Chapter 4 and 5 and identify the key features of each CEFR level. Though not suggested in the Manual, the outcome of the first session was a set of posters with the key features of each CEFR level. The idea came from an ETS TOEFL alignment project, which two of the colleagues who took part in that project found useful. Creating posters would give the participants a concrete purpose for analyzing the CEFR descriptors, it was felt, rather than simply discussing them. It would also help them clarify the differences between the levels. The Manual gave limited guidance on the focus or direction of the discussions.

4.2.2.2 Session 2 – Update

This session took place after the summer holiday and served as a reminder to refresh memories by recapping what was done in the first session and present the plan for the upcoming academic year. Project members went over the posters they had prepared in the first session and were given another opportunity to discuss the levels in relation to their own context.

4.2.2.3 Session 3 – Berlin update and article discussion

As mentioned in 4.2.2.1, two of the project members had taken part in an ETS (Educational Testing Service) CEFR linking project in Berlin and they gave an update of what they had done. The aim was to help members better understand what standard

setting involves and how CEFR levels are used in the project. In the same session Brian North's article entitled "The Common European Framework of Reference: Development, Theoretical and Practical Issues" (2006) was discussed to give background information on the CEFR and its levels.

4.2.2.4 Session 4 – Standard setting simulation

The two project members who took part in the TOEFL standard setting study conducted a session that provided a simulation of how CEFR scales were used in making judgments about test items. At the end of this session, a definition of a least able B2 candidate for reading, the basis of the reading standard setting method used, was produced to be used in the reading standard setting during the standardisation stage. This session provided a great opportunity to be proactive about the actual standard setting, addressing questions and issues regarding standard setting procedures.

4.2.2.5 Session 5 – Overview of the study on the construction of the CEFR descriptors

There was an informal evaluation discussion at the end of each familiarisation session. A request from almost all of the project members to learn more about the CEFR, and how the scales were constructed, came after the third one. The project leader and the researcher felt that it was clear from the discussion that members were intimidated by the perceived complexity of the scales, and decided to have another article discussion session to help participants have a good understanding of the CEFR. Prior to this session, the participants were sent an article by North and Schneider (1998) on how the CEFR scales were constructed and were asked to read it. The session started with some input on Rasch scaling so that the participants could understand the methodology used to construct the scales. Then, a review of the stages of the CEFR scaling study, as

described in the pre-session article, was undertaken. Throughout the session, there was an open discussion where participants asked questions and raised issues regarding how the scales were developed. The intention here was to inform participants about the origin of the scaled descriptors.

4.2.2.6 Session 6 – In-depth analysis of CEFR descriptors

This session aimed to further analyse the CEFR descriptors and scales to enhance participant familiarity with the CEFR levels. The project members mainly worked with the Illustrative scales or descriptors. Separate scales are available for each receptive and productive skill, and the skills scales are further broken down into scales for communicative activities, strategies, and language competences (Council of Europe, 2001). The purpose of working with the Illustrative scales was to identify which ones would be useful in the upcoming standardisation sessions as one or more of these scales would be utilized to assign levels to reading/listening items and written/spoken performances. The type of scale used has to be relevant to the context and tasks of the examination under study to achieve construct validity.

The participants also examined the overall reading, listening, writing and speaking scales with a view to detecting issues of parallelism and progression. For instance, the words 'understand' and 'recognize' both mean comprehension in some of the descriptors (Alderson, et al., 2004: 9). Another example is the use of the word 'infer'. Although making inferences may be needed at all levels even at A1 level, it is only mentioned in some of the CEFR levels. Such issues were potential problems for the project in that different participants could have different interpretations of the same scales due to the inconsistent use of certain terms. The participants also discussed the

Dutch CEF Construct Project report (ibid), which highlighted issues with the way the descriptors are phrased or constructed, giving suggestions to those who work with the CEFR. The project members then focused on the BUSEL Preparatory program syllabus and marking criteria to identify the overlaps and differences between the two systems; namely, BUSEL levels and CEFR levels. This session also contained a task, requiring participants to identify the CEFR levels of descriptors and rank order them. The participants worked with the descriptors individually so that they could see how well they understood the CEFR scales.

4.2.2.7 Session 7 – Meeting to discuss issues regarding embedding the CEFR into the BUSEL context (Invited speaker Prof. H.A.L. John DeJong)

Professor de Jong, who works closely with the CEFR, visited our school. During his visit six project members who were available at the time had a meeting with him where they had a chance to exchange views on the uses of the CEFR and how the familiarisation stage of a linking project should be handled. The outcome of this meeting was a list of suggestions regarding the familiarisation stage. The use of quizzes to monitor familiarity of the project members was one of the suggestions implemented from that point onward. Although this session was the last of the familiarisation stage, the quizzes were used as an extension of this stage administered at the beginning of the specification and standardisation stages.

4.2.3 Familiarisation stage research procedures

As stated in the preceding chapter, the aim of the familiarisation stage is to ensure that the participants of the linking process gain an in-depth familiarity with the CEFR and its descriptors, a prerequisite to other stages. Therefore, for the case study, data were

collected on how well people had internalized the CEFR and its levels. Unless participants are familiar enough with the CEFR scales and levels to the extent that they can differentiate between the levels and apply the scales, then the validity of the standardisation regarding judge agreement would not be achievable. It appears that familiarisation relates to context and cognitive aspects of validity, in that, these aspects of validity involve contextual parameters such as task demands and language knowledge that forms a cognitive load on part of the test taker. Understanding the CEFR and its levels means understanding the underlying language competences and the circumstances in which expected language competences take place in the CEFR, on which users attempt to base their examinations through linking projects. The researcher aims to investigate whether familiarisation activities suggested by the Manual helps participants to consider aspects of test validity.

As outlined in Chapter 3, a questionnaire and two quizzes evaluated participant familiarity with the CEFR levels, and field notes were analysed to see what aspects of cognitive validity as described by Weir (2005a) the participants referred to while trying to understand the CEFR. For instance, did they give importance to the kind of thought processes or the cognitive load required to answer an item? Furthermore, the rank-ordering tasks, as described in section 3.7.1.1.d of Chapter 3, were also analysed statistically using many-facet Rasch to further investigate familiarity. The questionnaire administered at the end of the first familiarisation session aimed to find out participants' reaction as to whether the activities suggested in the Manual (Council of Europe, 2003) achieved their purpose and helped them become familiar with the CEFR.

Two quizzes given out at the end of familiarisation measured the familiarity level of the participants with the CEFR scales: one in February 2008 prior to the writing standardisation; and the other in November 2008 prior to reading standardisation. As mentioned earlier familiarisation needs to be addressed prior to standardisation as well. Two additional quizzes were given out as part of the standardisation stage but they were not analysed since they served as familiarisation activities before standard setting.

4.2.4 Data analysis and findings concerning the familiarisation stage

4.2.4.1 Questionnaire

The questionnaire administered at the familiarisation stage consisted of five parts, as explained earlier in section 3.7.1.1 of the previous chapter. The aims of each part were as follows:

- Part 1: finding out about how familiar the participants were with the CEFR and its levels prior to the project;
- Part 2: investigating the effectiveness of the pre-session tasks in terms of guiding the participants in getting familiar with the CEFR levels;
- Part 3: evaluating the effectiveness of the session and the tasks used;
- Part 4: gathering feedback on the content material (CEFR Chapters 4 and 5);
- Part 5: identifying future demands and needs.

For each item on the questionnaire, frequency, weighted totals and means were calculated and each given a code except for items 4-6, which are True/False questions. Questions 1 and 2 are not reported as they require work-related information and are not relevant here. Questions 7 to 11 and 14 to 19 are also left out of the analysis given here as they were related to the delivery of the session, which is not relevant to the

discussion. The delivery of the session, which might impact on understanding of the CEFR, was successful with a mean range of 3.58 to 4.66. The complete questionnaire results, with 12 respondents, can be found in Appendix 4A. The mean calculations of the items in the questionnaire are in Table 4.1.

Table 4.1 Familiarisation Questionnaire Results

Item	Code	Mean (N:12) Range 1 (low) to 5 (high)
I. Background Information		
3	Familiarity	2.4
4	T/F – CEFR document	12T
5	T/F – CEFR levels	1T-11F
6	T/F – linking studies	6T-6F
II. Pre-session Reading Tasks		
12	Content	3.25
13	Tasks	4.08
III. Session		
20	Tasks 1, 2, 3	4.16
21	Task 4	3.41
22	Task 5, 6	4.00
IV. Content		
23	Global scale	3.41
24	Self-assessment scale	3.58
25	Language progression	3.25
26	Context-free	2.91
27	Ambiguity	3.16
28	Own context	2.91
29	Widely-known levels	3.00
30	Chapter 4	3.25
31	Chapter 5	3.25
32	Rationale	2.83
V. Future demands and needs – an open ended section		
- More time needed to work with the descriptors - Need to look at actual samples to conceptualise the levels		

Part 1 of the questionnaire, viz. background information in Table 4.1, shows that whereas almost everyone in the group knew the purpose of the CEFR and how many levels it consisted of, only half of the group had some knowledge about the linkage studies of some well-known exams to the CEFR.

Questions 12, 13, 20 and 21 were the additional tasks to those suggested in the Manual. The figures on these tasks, with means ranging from 3.25 to 4.16, reveal that the tasks contributed in some measure to the participants' understanding of the CEFR and its levels. Questions 12 and 13 targeted investigating the effectiveness of the pre-reading tasks in terms of guiding the participants in getting familiar with the CEFR levels. The results revealed that with a mean of 3.25 out of 5, the participants thought that the CEFR Chapters 4 and 5 were moderately easy to comprehend. However, the weighted total for this question (See Appendix 4A for the weighted totals) showed that responses leaned towards the 'not sure' band. This result is also backed up by questions 30 and 31, which also aimed at evaluating the content material viz. CEFR Chapters 4 and 5. These chapters are significant to understanding the cognitive requirements of each CEFR level as Chapter 4 deals with the context of language use; i.e. themes, tasks, purposes, activities and strategies (parameters of context validity) and Chapter 5 explores the cognitive processes involved in language use and text types (parameters of cognitive validity). Understanding the content material in these chapters is what allows judgments to be made about an exam or its items and tasks in a linking study. The participants are required to say whether the exam in question, COPE in this case, is actually measuring the processes and competences required by a given CEFR level, which is related to investigating the context and cognitive aspects of validity. In terms of assimilating the progression of the CEFR levels by identifying the key features of each CEFR level (Question 13), with a mean of 4.08, the participants expressed that the tasks were helpful, fostering aspects of context and cognitive validity. Among the six tasks used in the session, Task 4 (question 21), involving relating the CEFR levels to the BUSEL context in terms of how levels are realised, received the lowest rating.

Questions 23 to 32 aimed at evaluating the content material, that is the CEFR, its levels and use, received the lowest ratings relative to the other parts of the questionnaire, with means ranging from 2.83 to 3.58. Not all the participants were convinced that the CEFR could be used in any context and they found it difficult to relate the levels to their own context. This might be explained in a number of ways. Firstly, the participants may not have become familiar enough with the CEFR descriptors to form the link, or they might have also had problems reconciling the two systems, CEFR and BUSEL levels, as they are perceived to progress in different ways. For instance, making inferences comes into play at B2 onwards (B2 at the time seemed to correspond to our advanced/highest level) whereas in the BUSEL context, students learn, but are not tested on, making inferences at pre-intermediate level. A further explanation might be that without actual reading or writing samples, the participants may have had difficulty conceptualizing the requirements of CEFR levels, indicated by a couple of the participants in the 'Comments and suggestions' section of the questionnaire, saying that they needed to look at the scales and the descriptors in more detail. The last explanation appears to be the more plausible one as unless participants work with real samples and items, it is difficult to conceptualize the CEFR levels, as will be evident in the following stages of the process. From a research perspective, data regarding questions 23 to 32 showed that the linking process did not provide participants with activities that enabled them to understand the CEFR levels with respect to the context and cognitive aspects of validity in the early stages, i.e. familiarisation.

The data from the questionnaire in terms of the content material, CEFR, suggested that more sessions would be required to help participants better familiarize themselves with the CEFR and its scales. It also suggested that even though the tasks carried out in the

familiarisation, as suggested by the Manual, helped the group understand the CEFR levels to a certain degree, they were not sufficient enough to proceed in the process without applying the scales while examining sample exam items or performances. In other words, CEFR familiarisation activities lack in bringing out the demands of a given CEFR level, i.e. parameters of context and cognitive aspects of validity, in relation to items of an examination and how these items reflect the given CEFR level.

4.2.4.2 Field notes

The field notes presented in this section came from the first familiarisation session, as explained in section 3.7.1.1. The data was analysed for all aspects of validity viz. test taker characteristics, cognitive, context, scoring, criterion-related and consequential validity. Each parameter of a validity aspect as described by Weir (2005a) formed a code (See section 3.8.3 for details on the analysis of field notes). However, only two, context, cognitive and scoring validity, as presented in Table 4.2 occurred regularly in the analysis. The 'Themes' column in the table indicates the aspect of validity and the 'Descriptions' column shows the parameters of that aspect of validity that came up in the analysis. The 'Frequencies' column presents the number of times each parameter was raised by the participants throughout the familiarisation stage. As presented in Table 4.2 most of the discussion centred around the criteria, that is the CEFR scales and levels in this context, which is a significant element of scoring validity. This is an expected outcome as the familiarisation session aimed at working with the CEFR scales and thus getting familiar with them. It was the participants' task to analyse the descriptors in depth.

Table 4.2 Frequency of the Themes Occurring in Field Notes - Familiarisation

Themes	Descriptions	Frequencies
Context validity	Task design	9
	Task demands	2
	Language knowledge	7
Cognitive validity	Language knowledge	7
Scoring validity	Criteria	45

Providing evidence for scoring validity requires providing data on how well the judges know and can use the CEFR scales, which act as the criteria in this case. Judges need to be able to differentiate between levels clearly and until they reach this ultimate goal, judges are required to work with the criteria. Whether this goal can be realistically achieved or not by the end of the familiarisation session, as suggested by the Manual (Council of Europe, 2003), is addressed in the next section through the analysis of the rank ordering and quiz data.

4.2.4.3 Statistics (Analysis of rank ordering tasks and quizzes)

The analyses of the Manual rank ordering tasks, activities d and e (Council of Europe, 2003: 25), used in the familiarisation stage and the locally prepared quizzes on the CEFR writing and reading descriptive schemes aim to provide evidence on the levels of familiarity and consistency of the project members while working with the CEFR descriptors, as explained in detail in section 3.7.1.1 of the previous chapter. Knowing the CEFR scales and being able to use them consistently are associated with scoring validity, as explained above. The analyses of the rank ordering tasks used in the familiarisation stage are presented in Table 4.3. Consistency and agreement among judges are reported using Cronbach alpha, Pearson correlations, and intraclass correlation coefficient (ICC) as well as many-facet Rasch.

Table 4.3 Agreement and Consistency of Judges - Familiarisation

	Global	Read.	Listen.	Writ.	Qz1 Feb	Qz2 Nov
Alpha	.9935	.9970	.9974	.9977	.9734	.9695
ICC	.9935	.9970	.9974	.9977	.9734	.9695
Pear. Corr.	1.000	1.000	.9977	1.000	.7950	.7950
Mean Infit	.70	.76	.85	.62	.83	.92
Reliability	.00	.00	.00	.00	.43	.51

Alpha, ICC and Pearson correlation indices show that the judges performed successfully in the familiarisation tasks. However, the judges were less successful in the quizzes although considering the sample size, 10 items, the correlation coefficients are high. This might be due to the nature of the quizzes, in that, unlike the rank ordering tasks that required the judges to put the given descriptors in order of level, the quizzes asked them to recognise the CEFR levels of the given descriptors by recalling key features of each level and skill. This latter task may be considered more challenging.

FACETS outcomes in Table 4.3 where the mean infit statistics and reliability are reported for the judges are difficult to interpret. In analysing rater performance, an infit between 0.4 and 1.2 is considered reasonable by Linacre and Wright (1994). In this case, the mean infit values for all tasks are within this range. In terms of reliability, the judgement process was highly reliable with a value of .00. A Rasch reliability index does not indicate the degree of agreement between raters but how far they differ in terms of severity (McNamara, 1996). Therefore, the reliability index needs to be low for raters (Linacre, 2007). Here, although the mean fit statistics are within the acceptable range, analysis of each judge shows that for the familiarisation tasks there were several inconsistent judges (Table 4.4). However, this contradicts the raw data where these judges seem to have misplaced only one or two descriptors. It might be concluded that as the raw data was almost perfect, the FACETS program reflected slight drifts as big

deviations in the analysis. In all these tasks, the judges had logit scale values of .00, which confirms they were neither severe nor lenient (See accompanying CD Folder 1 Appendix 3C for the All Facet Vertical Rulers). This was again reflected in the reliability indices. However, the situation was different for the quizzes. The judges were only moderately successful, with a Rasch reliability of .43 and .51 for two quizzes; and there was one inconsistent judge as shown in Table 4.4. As mentioned previously, the nature of the task in the quizzes was different from that of the rank ordering tasks. The locally designed quizzes require knowing the features of each CEFR level. Institutional restrictions, at times there were long intervals between the sessions, might have made it difficult for the participants to activate their knowledge on the CEFR after long periods of time not working with the descriptors.

Table 4.4 Inconsistent Judges – Familiarisation (N=15)

Global		Reading		Writing		Quiz 1 Feb		Quiz 2 Nov	
Rater	Infit	Rater	Infit	Rater	Infit	Rater	Infit	Rater	Infit
1	2.45	1	2.45	9	1.81			2	1.50
2	2.45	4	2.14	14	1.81				
		14	2.14						

4.2.5 Familiarisation stage summary of research findings

The responses to the questionnaire showed that the familiarisation stage activities suggested by the Manual are not enough to bring out aspects of validity, context, cognitive and scoring in particular that are of profound importance to understanding the CEFR, its levels and scales in depth. Parameters of context and cognitive validity such as the purpose of a task and its linguistic demands are key to understanding the level set through an examination and a given CEFR level, which enables any judge to use the CEFR scales consistently, thus enhancing scoring validity. However, the fact that only linguistic demands come into play in the CEFR in terms of context and cognitive

validity is a concern as most other parameters of these aspects of validity such as cognitive processing strategies are not highlighted in the CEFR.

The analyses of the field notes again revealed the link between familiarisation and scoring validity because the judges mostly referred to features of the CEFR scales and levels throughout the familiarisation activities. Aspects of cognitive and context validity were also considered during familiarisation but to a limited extent.

The statistical analysis of the familiarisation activities suggested that the locally prepared quizzes were a more difficult but rigorous form of training as the quizzes required a deeper knowledge of the CEFR. The tasks given in the Manual, although focusing on scoring validity through rating difficulty levels, might not be the most effective activities to be used to train the participants of a linking study.

Overall, the researching the familiarisation stage of the linking process suggests that the strongest aspect of validation brought out at this stage is the scoring aspect of validity. Scoring validity deals with analysis of given criteria or 'the rating scale', 'the rating procedures (training, standardisation, rating conditions, rating, moderation and statistical analysis)', 'raters' and "grading and awarding". As far as these aspects of scoring validity are concerned, the familiarisation stage entails analysing the criteria – the CEFR scales and levels – to be used, training the raters – the judges or the project members in this context – and to a certain degree statistical analysis, since how well the raters mastered the scales is analysed statistically. However, it should be noted that the findings at this stage also demonstrated, to a certain degree, the strong relation between context, cognitive and scoring validity. In fact, context and cognitive validity are an

integral part of scoring validity. Without sufficient knowledge of the context and cognitive aspects of a test, scoring validity cannot take place because it is the context parameters and cognitive requirements of a task that help determine or design a set of criteria. The criteria have to reflect these parameters closely. Moreover, it is the knowledge of these parameters that enables raters to use the criteria, in this case the CEFR scales, successfully.

Regarding how reading and writing are tackled at the familiarisation stage, these skills are distinguished in so far as the tasks are separated for each skill, but no consistent discussion arose on the different ways CEFR linking approaches these two skills at the familiarisation stage.

With respect to the participants, extending the familiarisation stage was an appropriate decision as already indicated in Chapter 3 Research Design and section 4.2.2 of this chapter. By the end of this stage, it could easily be observed that all the participants started to share a common understanding of the CEFR levels and scales. At the end of every familiarisation session, time was allocated for comments, questions and evaluation of the sessions. After the last session, the participants reported that, feeling comfortable to work with the CEFR, they were ready to move on to the other stages of the linking, although they had concerns about what was ahead of them.

4.3 Specification stage

4.3.1 The Manual approach

The specification stage aims to ascertain whether an examination is designed and produced based on good practice as outline in section 3.7.1.2. The exam coverage is

reported in terms of the categories presented in CEFR Chapter 4 "Language Use and the Language Learner" and Chapter 5 "The user/learner's competences". This is considered to be a qualitative way of providing evidence through "content-based arguments" (Council of Europe, 2003: 2) such as themes covered and skills tested.

The specification process is composed of two phases: a general description of a chosen exam and a detailed description of this exam. The general description involves a global analysis by filling in Form A1, which deals with the aim of the exam, domains involved, communicative activities tested, duration, test tasks, information provided for test takers and teachers, and reporting scores (Council of Europe, 2003). The detailed description of the exam entails filling in Forms A8 – A22 by giving further details of each sub-test as regards Communicative Language Activities (CEF Chapter 4) and Aspects of Communicative Language Competence (CEF Chapter 5). The results of the descriptions are presented graphically demonstrating the exam coverage in relation to CEFR levels. The end-product of the specifications stage is a report that makes an examination more transparent to the test takers and the users of results (ibid).

4.3.2 Overview of the specification stage of the project

As stated above, this stage involved filling in the forms (A1 to A23) provided in the Manual (Council of Europe, 2003), some requiring a general description of the exam, others requiring a detailed description of each paper. The forms relevant to this research are given in Table 4.5. The Manual recommends that the completion of the forms is undertaken by the team responsible for the exam under study through a discussion or individually followed by a discussion (ibid). In the COPE linking project, a session including ten local participants and an invited participant Prof. Ozcan Demiral, the

Turkish delegate of the Council of Europe Language Policy Division and the European Language Portfolio National Coordinator was held. Four members of the COPE team were involved in the project and it was felt that involving the rest of the project group as well as an external expert could result in a more thorough analysis of the examination.

Table 4.5 Overview of the Specification Forms Relevant to the Research

FORMS		FOCUS
General description of the exam	A1	The general exam description
	A2	Test development
	A3	Marking
	A4	Grading
	A5	Reporting results
	A6	Data analysis
	A7	Rationale for decisions
Detailed description of the exam	A8	Impression of overall examination level
	A10	Reading comprehension
	A14	Written production
	A19	Aspects of language competence in reception
	A21	Aspects of language competence in production
	A23	Graphic profile of the relationship of the examination to the CEF levels

4.3.2.1 Prior to the session

Forms A1-A8, aiming at a general description of the exam, were filled in by the project leader and the researcher as some of the detail required was not known by the other group members. The descriptive sections of the other forms relevant for the COPE exam (A10, A14, A19 and A21) were again filled in by the project leader and the researcher but CEFR levels were not assigned so that the level allocation could be carried out through a group discussion.

4.3.2.2 During the session

The Manual (Council of Europe, 2003) suggests familiarisation activities be repeated prior to both specification and standardisation stages. In this case, the familiarisation

stage was limited to looking at the reading comprehension and written production descriptors, and discussing them in relation to the COPE exam.

The specification session opened with a discussion to clarify the terminology in the CEFR forms as some terms were considered to be ambiguous by the group. The group had to have a shared understanding in terms of how the CEFR describes its action-oriented approach to language learning. Once the terminology was clarified, participants were asked to look at the completed versions of the forms A10 (reading comprehension) and A11 (written production) in groups of three. Their task was to check whether they agreed with the completions of the project leader and the researcher by annotating their copy as to what additions, deletions or changes they would suggest. They were then asked to assign the required CEFR levels, for which they had access to a sample exam from the COPE item bank so that they could analyse the reading and writing tasks used in COPE and the level of reading texts as well as the expected output texts for writing. Participants were also given COPE test specifications and the CEFR document itself should they need more information about the purpose of each COPE section and all CEFR scales.

High levels of agreement were observed among the judges with a 100% agreement (all the judges assigned B2) on the level of written production and a 90% agreement (9 out of 10 judges assigned B2) on the level for reading comprehension. The justifications made for the levels were a group effort and recorded by the project leaders. Time constraints and the complexity of the forms for aspects of communicative language competence (A19 and A21) for reception and production meant that these would be filled in by the group members individually outside the session as the institution could

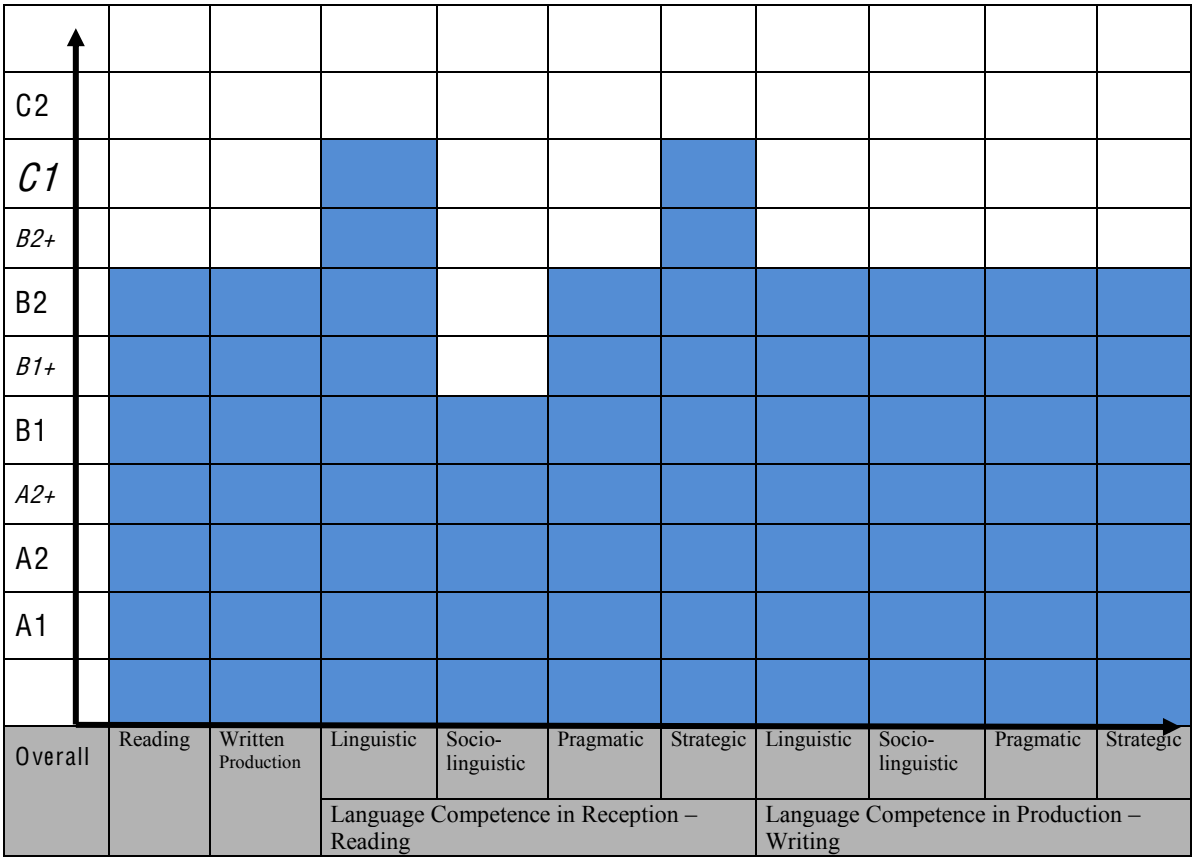
not accommodate another session due to the needs of daily operations and also because the following stages of the process had already been scheduled with externals experts invited on set dates.

4.3.2.3 After the specification session

After the specification session, the participants were sent the forms for aspects of communicative language competence in reading and writing (A19 and A21) to be filled in individually by looking at the relevant documents and assigning levels to language competence requirements of the COPE. When the forms had been returned, the project leader and the researcher collated and analysed them. The results showed low agreement (20% to 30% for different sections) in terms of the CEFR levels assigned and the descriptions made about the COPE examination. Moreover, when contacted to investigate how they filled in the specification forms, six out of ten indicated that they were not clear about how to fill them in, that is, they had the target group in mind not the test itself. In other words, while filling in the forms, the participants focused on the skills and language levels of the typical BUSEL exit level students rather than what the COPE examination aims to measure. The language competence forms for reception and production were collated and the level suggested by the majority of the group was included (six out of ten were in agreement). The forms were later sent back to the members again for a final check and approval. However, there was a low return because, as four members stated to the researcher when asked why they could not return the forms, that the questions in the forms were complex and they could not be sure about what COPE aimed to measure, for instance, in terms of socio-linguistic competence. Moreover, it seemed that judgment outside of a formal session could not be relied upon because of a lack of common understanding over terminology and what

was required of them because such problems were not experienced for the writing, reading and listening forms, which were completed in a session. The problem with the language competence forms could only be resolved by holding a further specification session to look at these two forms again as the session held for reading and writing were useful. Due to constraints resulting from the fixed timetables of the project members, another formal specification session could not be held. Time had already been set for the standardisation stage and, as it involved experts coming in from abroad, the date could not be changed and used for specifications instead. Therefore, the project leader and the researcher completed the language competence forms for reception and production themselves by referring to those forms sent by only three project members. Finally, a graphic representation of the specification stage results, which can be seen in Figure 4.1, was drawn.

Figure 4.1 Graphical representation of the COPE levels in relation to the CEFR



4.3.3 Specification stage research procedures

As mentioned in the previous section and in Chapter 3, some of the problems arising from the specification stage stemmed from participants' lack of understanding of the specification forms, and others were logistical. Therefore, the project leader and the researcher decided to focus on the data that came from three most experienced members of the team who were heavily involved in testing and who had experience of writing test specifications and were particularly familiar with the COPE exam. When the forms they filled in were analysed, it could be seen that they had detailed justifications for the statements they put in the forms with reference to the COPE test specifications and the CEFR. Analysis of their forms showed that they were also in agreement on the way they had completed the forms, although the form filling had been done individually. Other members of the group could also have contributed to the form completions if they had had the chance to sit in a formal session and discuss the problems they encountered completing the forms with one another. However, this was not possible in practice.

In order to research the specification stage, a group interview was conducted with the three people who best contributed to the completion of the specification forms. They were asked to look back at the specification forms to remind themselves of the type of questions they had to tackle while filling in the forms. Prior to the interview, they each were given a copy of the interview questions that arose from the research questions (See Appendix 4C for the specification stage interview questions).

4.3.4 Data analysis

The group interview was fully transcribed by adopting a very simple transcription scheme (See Appendix 3E for the interview coding scheme). Similar to Papageorgiou

(2007b), the interest was primarily in what the panelists said in response to the questions, rather than the nature of the interaction during the group interview. In other words, the researcher was more interested in the 'what' rather than the 'how' of participants' responses (Holstein & Gubrium, 2004: 142).

After transcription, the text was checked against the tape again to ensure accuracy. Once the quality was ensured, the data were coded manually as the amount of data to work with was manageable. The approach adopted for coding, as explained in detail in section 3.7.1.2a, was deductive (Creswell, 1994) in which key words and phrases in the research questions and the sub-questions were used as a starting point. Then, codes related to each sub-question were brainstormed using the aspects of validity explored by Weir in his framework. Throughout deductive coding, other codes also emerged and were added to the coding scheme. The researcher did check-coding herself and achieved a high level of code-recode consistency with 2 minor additions to the tapescript. It should be noted here that data regarding some of the codes were not found in the analysis but kept in the coding scheme to reflect the missing aspects of validity in the specification stage.

4.3.5 Research findings concerning the specification stage

The results of the group interview are discussed in three sections below related to each research question. The first section (4.3.5.1) examines how far the specification stage contributed to the validity of the COPE exam as a whole. The second section (4.3.5.2) explores whether the linking process has contributed to the validity of the reading and writing papers in a similar way. The final section (4.3.5.3) presents findings related to the impact of the specification stage on the level of the COPE examination.

In the following sections, the interview data are summarised in tables concering each aspect of validity. In data analysis, the interviewees were allocated numbers as I1, I2 and I3. The letter codes Y (yes) and N (no) were used to indicate whether they agreed with the issues explored by other interviewees; Y indicating agreement and N indicating disaagreement. The symbols (+) and (-) were used to show two dimensions of the code. For example, if the specification forms were effective in making the interviewees consider a parameter of context validity, that parameter was assigned a (+), if it did not, it was given a (-). In addition, the frequency of the codes mentioned related to every parameter was also indicated under the frequency column.

4.3.5.1 The contribution of the specification stage to the validity of the exam

(a) Test taker

Codes regarding the test taker did not emerge in the analysis of the group interview, which might suggest two reasons. First, the interview questions may not have focused on the test taker, failing to draw the attention of the interviewees to this aspect of validity. However, the group interview took place in the form of a semi-guided interview where the interviewees were given the questions and prompts for each question to guide them in advance. One of the prompts was the test taker and the interviewees did not talk about it. They might have focused on the main interview questions and used the prompts they thought were relevant to what they wanted to say. Second, it might suggest that the characteristics of the test taker were not taken into consideration at the specification stage.

(b) Context validity

Context validity relates to task design, task demands and the administration of the exam. as indicated in Table 4.6. In terms of task design, the process of filling in the specification forms did not make the interviewees question the design of the COPE exam. The questions in the specification forms related to design were mechanical and did not require thorough consideration of the exam design. Among parameters of task demand specified by Weir (2005a) such as the nature of information, text length, and content knowledge, the only aspect of context validity regarding task demands that the interviewees referred to was the linguistic demands of the exam. They mentioned that the linguistic competence forms made them think about the exam and the expectations of the test takers. I1 was in fact critical of the thought process she went through while filling in those forms.

> "*I think to fill that properly, again what you said, you'd have to spend a lot of time really analyzing the lexis, the structures of the actual, I'm talking about reading tests. Those were the most challenging and I think thought-provoking. For reading, we had linguistic competence that was valid but sociolinguistic competence one didn't seem to relate to our reading test as far as I can remember. The pragmatic competence and the strategic competence, how can you answer that? That's quite a difficult one to think about what strategies we are expecting the test takers to use when they deal with the COPE reading paper. So those I found those really useful"* (S: 323-330).

Table 4.6 Context Validity – Specification Stage Interview

Themes	Descriptions	Frequency	I1	I2	I3
Contribution to context validity	Task design (+)	2	Y	Y	Y
	Task design (-)	9	Y	Y	Y
	Task demands (+)	4	Y	Y	Y
	Task demands (-)	4	Y	Y	Y
	Test administration (+)	-	-	-	-
	Test administration (-)	4	Y	Y	Y

However, the interviewees repeatedly emphasized that although the forms made them think about the demands of the task, filling in the forms did not contribute much to their understanding of the exam and that analysis of texts and tasks could only be effective through group review of items and discussions.

With respect to test administration I2 pointed out that the fact that the forms did not draw users' attention to the conditions under which the exam was administered was an important lack in the forms. He emphasized the importance of administration by saying *"If you have improper administration then you're gonna have improper results no matter how good your grading is or how well you've designed your exam"* (S: 139-140).

(c) Cognitive validity

The frequencies in Table 4.7 suggest that the issue with respect to cognitive validity was limited to the cognitive load in terms of the language knowledge required to complete a task. The arguments put forward by the interviewees supporting this finding mostly overlapped with those for task demand under context validity since language knowledge is a feature common to both. The specification stage helps users to think "more categorically" (S: 290), in that, the forms provide users with areas to focus on while exploring cognitive validity of an examination; however, similar to context validity, the

interviewees pointed to the importance of having a group examine tasks and have a discussion on them that it is crucial in understanding the language knowledge required by the tasks in the exam. As was the case in the familiarisation stage, language knowledge is emphasised in the CEFR over cognitive processing, which seems to be a lack on part of the descriptors.

Table 4.7 Cognitive Validity – Specification Interview

Themes	Descriptions	Frequency	I1	I2	I3
Cognitive validity	Language knowledge (+)	8	Y	Y	Y
	Language knowledge (-)	6	Y	Y	Y

(d) Scoring validity

Scoring validity seems to be the main weakness of the specification forms as only once throughout the interview, an interviewee mentioned the "proper answer key", "proper criteria" and "statistical data" (S: 339-340).

Table 4.8 Scoring Validity – Specification Interview

Themes	Descriptions	Frequency	I1	I2	I3
Scoring validity	Item analysis (+)	1	Y	Y	Y
	Item analysis (-)	-	-	-	-
	Criteria (+)	1	Y	Y	Y
	Criteria (-)	-	-	-	-

(e) Consequential validity and Criterion-related validity

Consequential and criterion-related aspects of validity are presented together here and in Table 4.9 because the interviewees talked about these two aspects in relation with one another regarding how test data is used after an examination is administered. The interviewees indicated that the specification forms "take (them) beyond the marking and scoring into the data and the data interpretation stages" (S: 347-348). They indicated

that although it is mostly a box ticking exercise in the specification forms, it still helps the users become aware of the areas they are supposed to focus on. As the frequencies in Table 4.9 show, the emphasis was on the consequential aspect of validity more than the criterion-related aspect.

Table 4.9 Consequential and criterion-related validity – Specification Interview

Themes	Descriptions	Frequency	I1	I2	I3
Consequential validity	Washback (+) Waschback (-)	7 -	Y -	Y -	Y -
	Effect on individual (+) Effect on individual (-)	4 -	Y -	Y -	Y -
Criterion-related validity	Other tests (+) Other tests (-)	1 -	Y -	Y -	Y -
	Future performance (+) Future performance (-)	1 -	Y -	Y -	Y -

4.3.5.2 The contribution of the specification stage to productive and receptive papers

Codes related to reading and writing skills that emerged from the analysis of the group interview did not yield any information regarding the contribution of the specification stage to productive and receptive skills separately. The interviewees referred to the reading and writing papers while they were giving examples for issues relevant to both papers. However, one of the interviewees (I1) mentioned that the language competence forms for reading were very useful. For instance, she stated that it was "quite a difficult one to think about what strategies we are expecting the test takers to use when they deal with the COPE reading paper" (S: 329-330). The reason why she thought the language competence forms for reading were particularly difficult to fill in might be because the cognitive requirements of receptive skills cannot be observed whereas some of the requirements for productive skills such as initiating, organising etc. are observable.

4.3.5.3 The implications of the specification stage on the level of the exam

As regards the level of the exam, the analysis of the group interview centred around three aspects of level; intended level, understanding of the level, and increasing the standards as presented in Table 4.10. The interviewees found the process of filling in the specification forms "reassuring" because their "gut intuitive feeling (was) backed up" (I1). When the graphical representation of the level of the exam was filled in for reception, production and language competence, the end product was an uneven profile. This made them start questioning the unevenness and in the end they came up with justifications for the uneven profile which they had never thought of openly before in line with the relevance to their context. For example, sociolinguistic competence for the COPE examination came out to be at B1 level and the rationale for this was that sociolinguistic ability was not a focus on the BUSEL syllabus. Thus, students were not expected to have a high level of sociolinguistic ability, but this had not been thought or verbalized explicitly before. The interviewees saw the filling in of the forms as a positive experience, which helped them understand the COPE exam better.

Table 4.10 Institutional implications – Specification Interview

Themes	Descriptions	Frequency	I1	I2	I3
The level of the exam	Intended level	3	Y	Y	Y
	Understanding of the level	4	Y	Y	Y
	Increasing the standards	2	N	Y	Y

As regards the issue of increasing the level of the exam, this meant using the forms for a purpose other than what they were intended for (I1). The forms were designed to define the level of a test with respect to the CEFR. However, I2 and I3 indicated that specification stage would not on its own help increase the standards but the categorization set in the specifications, that is, thinking not only in terms of reception

and production but also with respect to sociolinguistic competence, pragmatic competence etc. would help in a discussion on how to increase the level of a test. Whether the specification stage could help increase the standards of a test was the only area throughout the interview where one of the interviewees disagreed with the others.

4.3.6 Specification stage summary of research findings

The above analysis suggests that the contribution of the specification stage to the COPE examination is restricted to facilitating a deeper understanding of what the COPE tasks require test takers to do and of the level of the examination, to a certain degree. Specifically, in terms of the test taker, the analysis of the group interview suggests that the specification stage does not focus on the test taker characteristics.

With respect to context and cognitive validity, the interviewees specifically focused on the language competence forms for reception and production that made them carefully analyse the test and define what each task was measuring in terms of grammatical and lexical knowledge, pragmatic, sociolinguistic, strategic competence etc. These aspects of language competence questioned in forms A19 and A21 (aspects of language competence in reception and aspects of language competence in production) lead the interviewees to think what reading and writing tasks were actually measuring, which contributes to context and cognitive validity arguments of an examination. In other words, the process of filling in forms A19 and A21 lead to validity concerns among the interviewees, causing them to reconsider the competences measured in the COPE examination. However, the language ability levels being underspecified in the CEFR made the process of form completion quite challenging.

Regarding scoring validity, the analysis of the group interview suggests that the specification stage did not contribute to the scoring validity of the COPE examination. However, the interviewees indicated that form A5 draws attention to "how the test data is reported" and "how the data is used" (S: 341-348), fostering parameters of the consequential aspect of validity such as the effect of the exam on the individual (Weir, 2005a). With respect to criterion-related validity, the analysis of the group interviewed revealed that the specification forms recommend users to carry out criterion-related validity studies.

In general, the contribution of the specification stage to the validity of an examination under study seems to be in the form of an awareness raising activity, with some aspects of validity such as the cognitive aspect receiving more attention. The forms in fact draw attention to several aspects of validity such as "construct validity" (S: 96) through form A2 which requires users to specify which aspects of validity are estimated for the examination.

The process of completing the specification forms lead to a better understanding of the level of the examination and what it measures. The forces the users to question the level they are testing at and whether that level is desirable or not. O'Sullivan also reports a similar experience with the City and Guilds project, in that, he proposes that filling in the specification forms "forced the team to consider aspects of the tests not necessarily referred to directly in the re-written specifications" (2009a: 17).

Overall, the specification stage was of great help in some respects, i.e. analysis of language competence levels required to complete test tasks, whereas it had major shortcomings in some others, e.g. test takers and administration. Although the Manual (Council of Europe, 2003: 7, 33) suggests that a weak linkage claim can be made about an examination, no claims regarding such an alignment to the CEFR should be made at this stage. First of all, the activity undertaken at the specification stage is an awareness raising activity and seems to entail tasks similar to writing test specifications for an examination but with a narrow scope. In addition, to be able to say that even at this 'intuitive' level an examination has some kind of a link to the CEFR, evidence, even if it is only limited to qualitative data, should be provided as to how this stage was undertaken, ideally with a group of people. At least a figure or percentage should be reported regarding consensus among participants. The Manual seems to suggest that the completed forms and the graphical representation of levels are "sources of evidence to support claims" of linkage to the CEFR (ibid: 30). However, without any evidence on the quality of the process gone through in filling in the forms, they are unlikely to present any credibility. Moreover, linkage claims purely based on content coverage without standard setting remains meaningless and insignificant because standard setting is "the proper following of a prescribed, rational system of rules or procedures resulting in the assignment of a number to differentiate between two or more states or degrees of performance" (Cizek, 1993: 100). Following this definition, writing a test that reflects a certain CEFR level in terms of content does not provide you with a cut score or the number that differentiates between levels. In other words, one might argue that a test measures all aspects of the B2 level; however, how can one know how many correct questions on that test is required for a minimum B2 performance or a good B2 performance? Though not rationalized specifically, O'Sullivan also suggests that "any

such claim at this point is likely to be premature and possibly even meaningless" (2009a: 18).

The findings of the specification stage reflect the experience in a specific institution and may show variations for different institutions depending on how this stage is undertaken.

4.4 Standardisation stage

4.4.1 The Manual approach

The aim of the standardisation stage, as the name suggests, is to ensure that the CEFR levels are implemented consistently by the participants involved in the linking process (Council of Europe, 2003). The standardisation process comprises four stages:

- Familiarisation: So as to be thoroughly familiarized with the CEF levels, the participants are asked to again do the kinds of activities described in the familiarisation stage of the linking process.
- Training: The participants use the standardized exemplars, if available, to assess learner performance for productive skills or assess the difficulty of items for receptive skills in relation to CEF levels. They justify their judgments and acquire experience in relating performances or items to CEF levels. Reaching consensus is crucial for training purposes.
- Benchmarking performances: This is the application of the consensus reached at the training to the assessment of local test samples of institutions undertaking linking studies and involves activities similar to the ones carried out in the training. The outcome is a set of locally standardized items or performances.

- Standard – setting: It is the process of setting cut scores for the different sub-tests in an exam in relation to CEF levels. The initial judgments need to be confirmed with empirical evidence gathered from real administrations. It also involves investigation of internal and external validity.

4.4.2 Overview of the standardisation stage of the project – writing and reading

In this section, the sessions held for writing and reading respectively are described; first of all, in terms of the number of sessions undertaken and what they involved; secondly, with regard to the tasks carried out for familiarisation, training and standard setting parts of standardisation.

4.4.2.1 Writing

The intention of the writing standardisation was to link the COPE examination to the CEFR at B2 level as B2 seems to be considered adequate for academic study by most English medium universities around the world (PGMAC, 2012). This section contains information and data about the standardisation process relating to determining a cut-score, for which three standardisation sessions were held.

The first standardisation session for the writing paper took place between April, 25-26th, 2007 and involved looking at the six broad bands in the COPE criteria and the relationship between these bands and the CEFR. The highest score possible for the COPE writing paper is 30 and these points are spread among six broad bands as Very poor (1), Poor (2), Inadequate (3), Adequate (4), Good (5), and (6) Very good. The number of COPE samples used also limited in this session; only five samples were used. Of the 14 judges who took part in this session, 12 were COPE CEFR linking project

members, and two were external experts. One external expert was a PhD student at the time whose PhD was based on a CEFR linking project with Trinity College Examining Board in Britain. The other had been involved in CEFR linking projects over a number of years with the City and Guilds Examining Board, also in Britain. It was agreed that having an external perspective would be most beneficial for the project as it would allow for a more critical analysis and dialogue concerning both the objectives that are tested and the actual level of the COPE exam.

The second session held on September 28th, 2009 had to be done almost two years later. The initial project framework planned the process in parallel therefore the standardisation of the other skills were carried out in sequence and, it took 2 years for the writing sequence to come around. The second session involving 11 local judges aimed at working with a much broader number of samples (85 samples) and linking the COPE scores out of 30 with the CEFR scales at B2. As it would be very hard to discuss the level of 85 papers one by one with the whole group, samples were distributed amongst the participants, ensuring that each sample was double marked. Many-facet Rasch was used to calibrate the samples. Due to unreliable judgments made in the second session, explained in 4.4.4.3, a third session was needed.

The third session, held in two parts on November 19th and 26th, 2009, involved looking at 20 samples with individual marking, followed by a round of discussion with 11 judges again. A satisfactory link between the COPE writing scores out of 30 and the CEFR was established after this session.

(a) Writing familiarisation

As the group had already participated in an extensive familiarisation program, the pre-task set for the first session was for all participants to refamiliarise themselves with the following Illustrative Scales as they reflected the task used in the COPE: Overall Written Production; Reports and Essays. The external participants were also sent a COPE exam with the test specifications so that they would be familiar with the exam before the first standardisation session.

At the beginning of the first standardisation session, the CEFR writing quiz was given to all participants as a familiarisation activity, designed to ensure that all the judges had an in-depth understanding of the CEFR descriptors. The quiz was marked together as a group, followed by a detailed discussion of the descriptors. No statistical analysis was carried out on the familiarisation quiz at this stage of the project as previously clarified in 4.2.3.

Familiarisation activities were not repeated for the other two sessions, as suggested by the Manual, as the Manual suggestions and the locally prepared quiz was known to the participants and would lose its purpose when carried out a second time. However, familiarisation was handled through allocating plenty of time for and incorporating discussions of the CEFR writing scales into the training parts of standardisation.

(b) Writing training

Prior to the first writing standardisation session, as advocated in the Manual (Council of Europe, 2003), for the purpose of training the judges, a number of CEFR calibrated writing scripts were chosen. As the intention was to link the COPE to the CEFR at B2

level, samples illustrating B2 and C1 levels were chosen. Three of the samples were taken from Cambridge ESOL and one sample from IELTS. The Cambridge ESOL samples, although benchmarked to the CEFR, are very different from the essay task that the test takers encounter in the COPE exam and therefore they are difficult to relate to the academic context, in particular the short report format. For this reason, it was felt necessary to include an IELTS sample. Currently, IELTS has not provided samples that are empirically linked to the CEFR, but they do provide a guide which relates scores on IELTS writing paper to the CEFR descriptors (See Taylor, 2004). The samples used for the writing standardisation are given Table 4.11.

Table 4.11 Sample Writing Papers Used for Standardisation

Sample	Exam	Level
1	FCE	B2
2	CAE	C1
3	FCE	B2
4	IELTS	A sample graded at B2

For the second and third sessions, the same samples used in the first session with the addition of three CEFTrain samples were used. CEFTrain is a transnational initiative supported by the European Commission Comenius Programme to promote CEFR standards in teacher education. As well as providing general information about the CEFR, it consists of reading, writing, listening and speaking exemplars with guidance to the CEFR levels they belong to (www.ceftrain.net).

(c) Writing standard setting

The first standardisation session of local samples took place immediately after the standardisation training. The COPE writing paper has only one format, the essay. 5 samples were chosen for the first standardisation by the project leader and the

researcher from the standardisation documents for a live version of the COPE examination. A number of student scripts from the live exam were selected, then marked by the core group of COPE writing markers. As consensus on the allocated grades had already been reached by this core group of 18 people, these were considered to be the best samples to be used. Further discussion of this takes place under the empirical validation stage in this chapter.

The scripts for the second standardisation session were also chosen from a live exam and marked by the core group of COPE writing raters. 11 CEFR project members were each given a pack of 12 to 13 COPE writing samples to be marked using the CEFR and the packs were arranged to allow for double marking of each script. The data were analysed using many-faceted Rasch to calibrate the scripts so as to align COPE scores out of 30 onto the CEFR writing scale.

Session 3 included live COPE written samples first marked by the core group of COPE writing raters and then by 11 CEFR project members. 20 of these samples were used in the last standardisation session. The participants assigned CEFR levels to 11 out of 20 samples through whole group discussions. This was followed by individual marking of 9 samples to gather data on the reliability of the standardisation stage. The individual marking was followed by a round of discussion. After the discussion, the participants had a chance to reconsider their judgments. In addition to 9 COPE scripts, 2 Cambridge samples were included in order to investigate how well the judges followed the levels set through calibrated exemplars.

The method used for the writing standardisation was the Examinee Paper Selection method (Hambleton, et al., 2000), which is also known as the Benchmark method (Faggen, 1994). The judges were given the COPE written samples and asked to read the scripts and consider the performance in relation to the criteria in Table 5.8 of the Manual (Council of Europe, 2003: 81). The procedure followed in sessions 1 and 3 was to have as many rounds of judgements as necessary (3 rounds for session 1 and 2 rounds for session 3) followed by data analysis and discussion. After each round, the average scores allocated to each of the samples were shared with the judges who were then asked to justify their judgements. After a lengthy discussion, the judges were given the opportunity to reconsider their initial judgement if they so wished and further discussion was held at the end of round 2. In session 3, as the group gained more experience in the process, they were satisfied with 2 rounds.

4.4.2.2 Reading

The standardisation for the COPE reading paper required two different sessions. Each session is described separately and referred to as session 1 and session 2. The institutional aim of the standardisation sessions was to link the COPE Reading Paper to the CEFR at B2 level. Session 1 took place on June 30th, 2007 with 14 judges. 11 of the judges were from the BUSEL CEFR project group and 3 international experts in assessment and the CEFR also attended. The second session was held on November 26th, 2007. The participants for session 2 consisted of 10 of the CEFR project members.

(a) Reading familiarisation

By the time the first Reading Standardisation session was held, the project members had already spent one year working and becoming familiar with the CEFR. For this reason,

the familiarisation activity was set prior to the session. The group members were asked to re-familiarise themselves with the following Illustrative Scales that were relevant to the COPE reading tasks.

- Overall Reading Comprehension
- Reading for Orientation
- Reading for Information and Argument

Before the session the group members were also asked to look at the poster they had previously prepared describing the Least Able B2 Candidate during the standard setting simulation in one of the familiarisation sessions (Section 4.2.2) and to come to the session prepared to discuss and possibly modify the poster.

Prior to session 2, the judges were again asked to familiarise themselves with the relevant reading descriptors and the Least Able B2 Candidate Poster. At the beginning of the session, there was extensive discussion of what was meant by the Least Able Candidate and the rationale for defining such a person was clearly explained. When all the judges felt confident about the task in hand, the group was then ready to proceed to the next part.

(b) Reading training

As is the case with the training using performance samples, the objective of the training activities for the Reading standardisation was to ensure that the judges had a common understanding of the CEFR levels for reading and that they could later apply this understanding when it comes to relating the COPE items to the CEFR levels. As

presented in section 4.2.3 and above, two standardisation sessions were carried out for the reading paper.

The first session started with the analysis and assessment of CEFR calibrated reading tests taken from Cambridge ESOL and the Finnish Matriculation Examinations (Council of Europe DVD). Table 4.12 present the tests used for training purposes. The Finnish Matriculation sample proved to be the most effective as this exam was unknown to the majority of the participants. The general consensus regarding the Cambridge ESOL tests was that it was difficult to look at them objectively as the level was already known, for example, everyone associated First Certificate exams with B2 level and that strongly influenced any decision made.

Table 4.12 Sample Reading Tests Used for Session 1 Training

Item	Test	Level
1	Finnish Matriculation Examination	B2
2	CAE	C1
3	FCE	B2
4	FCE	B2

In the second session, for the reasons discussed above, it was agreed that only the Finnish Matriculation Samples would be used for training purposes. The Finnish samples provided included only 1 text of four multiple choice items, which reflects the COPE reading paper task type. This was not regarded as an issue as the participants had worked with a wider range of items in Session 1. Training in Session 3 involved looking at the DIALANG and CEFTrain samples.

(c) Reading standard setting

Before the actual standard setting commenced, a number of decisions were made regarding the procedures for rounding and adjusting the established cut scores. The standard setting methods – Yes/No method, a variant of the Angoff method and the modified Angoff method – used in this linking project do not yield exact cut scores and therefore, the results need to be rounded. Cizek and Bunch (2007) suggest two ways of rounding cut scores. The first one is rounding to the nearest score and the second one is rounding up as the first obtainable score does not quite meet the cut. For the purposes of this project, the more conservative method is preferred so as to be certain about the cut score for a B2 level candidate.

As the cut score is derived by calculating the mean of the participants' individual cut scores, the standard error (SE) of the mean is calculated and then this error estimate is used to adjust the overall cut score. The need to adjust cut scores arose from the fact that for most standard-setting procedures,

> "the cut score is an estimate, not in the sense of a population parameter, but a statistic that is subject to random fluctuation and that would differ to some extent in replications of the procedure under similar conditions, with a different (though equivalent) group of participants, and so on" (ibid: 300).

Prior to the standardisation, it was also necessary to discuss the least able B2 candidate profile and as a result, the poster referred to in 4.2.2.4 was extensively modified and used as the basis for making judgements on the items during standard setting.

For the first standard setting session, the project co-ordinators selected a complete COPE reading exam with the accompanying item analysis data. A variant of the Angoff Method (Council of Europe, 2003: 91), the Modified Angoff (in Cizek & Bunch, 2007: 83) and the Yes/No method by Impara and Plake (ibid: 88-89) were employed for the Reading paper standard setting. The decision to use three standard setting methods was made on the premises that different methods may yield different cut scores. The use of a variety of methods would help compare and confirm the cut score established as a result of different methods. However, as the session progressed it became apparent while discussing the justifications for the assigned levels that the judges were having problems in making decisions mainly due to the difficulty of defining and working with the least able B2 candidate profile. Their justifications stemmed from the actual CEFR B2 reading descriptor rather than the least able B2 descriptor. The statistical analysis of the data from the session is presented in section 4.4.4.3.

In the Reading standard-setting session 2, only two methods were used; Yes/No method (in Cizek & Bunch, 2007: 88-89) and the modified Angoff method (ibid: 83). The reason for this was that working with three different methods was time consuming and the time restrictions for session 2 necessitated reduction in the number of the methods used. While deciding which of the three methods would be dropped, the project leader and the researcher decided that working with two methods which both focused on the least able B2 candidate would help the judges, in that they would not have to change their mindsets depending on the method. In other words, these two methods both required conceptualizing a borderline B2 candidate performance whereas the variant of the Angoff method required judges to think of any B2 candidate while making judgments on reading items.

The judges were asked to analyse the reading items and fill in the given task sheet whereby they had to indicate the CEFR level of an item and note down the justification for their judgments. Two rounds of judgments were carried out. At the end of round 1, item analysis data was shared and a discussion held where the judges were asked to justify their judgments regarding the items. The second round gave an opportunity to change their judgments should they have wished.

4.4.3 Standardisation stage research procedures

Data were, first of all, collected through questionnaires given to project members (one for reading standardisation and one for writing standardisation) to evaluate the standardisation stage. Unless the sessions had been perceived as effective, then the resulting cut scores would not have been reliable. A session that has problems in terms of its delivery, the standard setting methods used or the application of these methods could distort the cut score established in that session. Although more than one standardisation session was held for each skill, the questionnaires were administered only after the first writing and reading sessions because subsequent sessions were replicates of the first ones, in that, the same methods, task sheets, items were used and same procedures were followed. Field notes were also kept by the researcher and the project leader in order to identify to what extent validity or test level issues were considered during the discussions. This stage of the linking process also involved statistical analysis of judgments, specifically: descriptive statistics; Cronbach alpha; ICC; and many-facet Rasch analysis in order to investigate the contribution of the standardisation stage to the validity of the COPE exam particularly with respect to scoring validity (RQ1d). In other words, the validity and reliability of the standard setting was investigated.

4.4.4 Data analysis and findings concerning the standardisation stage

This section presents the data analysis and the findings of the standardisation stage. Data and findings regarding each instrument used, viz. questionnaires, field notes and statistic analyses, are presented respectively, first for writing and then for reading, to allow for comparisons of the results for both skills.

4.4.4.1 Questionnaire

(a) Writing

10 project members and 2 external experts participated in the COPE first writing standardisation. The questionnaire administered at the end of the session comprised of four sections: Standard setting session; Calibrated samples; CEFR descriptors; and CEFR linking process (See Appendix 4D for the questionnaire results).

Part 1 of the questionnaire focused on the effectiveness of the standardisation session. Unless perceived effective, the results of the standardisation stage, that is, the recommended cut off score, could not be relied upon. It is clear from the questionnaire findings that participants found the standardisation session highly effective with a 100 percent agreement across all 16 questions, except for question 8 with 91.66 percent. The latter focused on whether the participants felt that they had equal opportunity to contribute their opinions. It appears that one participant was not happy in this respect but did not give a reason for it.

Part 2 of the questionnaire, with three questions, intended to find out whether the calibrated samples provided by the Council of Europe were useful and sufficient. As table 4.13 suggests, although the Cambridge samples helped the participants with the

training of the writing standardisation, the participants felt that they did not clearly reflect the level they were claimed to be at, nor were relevant for use in an academic context, with a weigthed total of 41.6 percent. A possible explanation of this could be that the tasks in these samples reflected the general English domain rather than the academic. A further interlinked reason is that the Cambridge writing tasks were very guided and therefore very little language was generated.

Table 4.13 Proportions of Responses in Two Categories – PART 2 Cambridge ESOL Samples

Questions	Strongly Agree / Agree	Disagree / Strongly Disagree	%
1 standardisation	100		100
2 level	58.3	41.6	100
3 academic	54.54	45.45	100

Part 3, which had 5 questions as seen in Table 4.14, aimed at investigating whether the participants found the CEFR descriptors workable and relevant to their context. These areas are within the domains of context, cognitive and scoring validity. Most of the participants (66.6%) found the Written Assessment Criteria Grid provided in the Manual (Council of Europe, 2003) to be used at the standardisation of the writing easy to use and indicated that especially the 'plus' (+) levels were useful in making judgments (91%). The majority of the group (75%) also agreed that the writing scale could be used in any context including the academic context. However, some of the participants (45.45%) did not quite agree that the grid covered all aspects of writing, the overall scale part in particular.

Table 4.14 Proportions of Responses in Two Categories – PART 3 CEFR Descriptors and the Assessment Scale

Question	Strongly Agree / Agree	Disagree / Strongly Disagree	%
1 criteria grid	66.6	33.3	100
2 + levels	91.6	8.3	100
3 context free	75	25	100
4 academic	75	25	100
5 aspects of writing	54.54	45.45	100

The purpose of Part 4 with five questions was to evaluate the contributions of the Writing paper standardisation to the COPE examination. Table 4.15 suggests that the process forced the participants (91.6%) to reconsider the level of COPE and helped them (100%) understand the level of the examination better in terms of what it measures. It also pointed to areas for revision if the level of the exam were to be increased (100%) and had a positive impact on the reliability (91.6%) and the validity (100%) of the COPE examination.

Table 4.15 Proportions of Responses in Two Categories – PART 4 CEFR Linking Process

Question	Strongly Agree / Agree	Disagree / Strongly Disagree	%
1 reconsidering level	91.6	8.3	100
2 understanding level	100		100
3 increasing standards	100		100
4 reliability	91.6	8.3	100
5 validity	100		100

(b) Reading

11 project members took part in the second session of the COPE Reading paper standardisation. Similar to the writing questionnaire, the reading questionnaire administered at the end of the session comprised of four sections viz. Standard setting

session, Calibrated Samples, CEFR descriptors and CEFR linking process (See Appendix 4E for the questionnaire results).

Part 1 of the questionnaire focused on the effectiveness of the standardisation session. Unless perceived effective, the results could not be relied upon. It can be said that the participants found the standardisation session relatively less effective that the writing with a strongly agree/agree categories proportion of minimum 75% across 16 questions except for questions 6 (58.33%), 7 (33.33%) and 11 (66.66%). Questions 6 and 7 were related to time allocations for doing the tasks and the discussions, which some of the participants found inadequate. The second session had to be done over two slots as the work could not be completed in the first one and thus a second one was arranged to finish it. Perhaps the participants were simply commenting on the fact that a second slot had to be arranged. Question 11 was related to the first round of discussions held after the participants had a chance to assign levels individually to the reading items. Some of the participants felt that the discussions did not help them make their judgments. This might be perfectly acceptable when participants are not convinced or influenced by justifications put forwards by others as to why they thought a certain items was at a particular CEFR level.

Part 2 of the questionnaire intended to find out whether the calibrated samples provided by the Council of Europe were useful and sufficient. Table 4.16 reveals that among the sample items provided, with a highly agree/agree categories proportion of 100%, the most useful in the training part of the standardisation came from the Finnish Matriculation Examination. The items from that same exam were also found to be relevant to the academic context. While the participants had doubts about the items

being representative of the levels claimed, the FCE items seemed to be the least suitable for the purposes of CEFR linking in BUSEL. Only 25% thought that the FCE items reflected the levels they claimed to be at. A possible reason could be the task types were perceived different from the COPE task types, in that it was difficult to relate them to the BUSEL context because 91.66% of the participants indicated that the FCE items did not reflect the academic context. The overall impression of the calibrated samples, particularly Cambridge ESOL samples, suggests that scoring validity might be an issue since these samples were used for training purposes before moving onto standard setting. In other words, training plays an important role in scoring validity of writing tests as it enables establishing the expected standards to the rater. Similarly, the calibrated CEFR exemplars are provided by the Council of Europe to help users better understand the levels and relate descriptors to actual performance. Unless users can clearly see this relation or link, they cannot set the expected standards.

Table 4.16 Proportions of Responses in Two Categories – Cambridge ESOL & Finnish Matriculation Exams

Question	Strongly Agree / Agree	Disagree / Strongly Disagree	%
1 FCE standardisation	75	25	100
2 CAE standardisation	83.33	16.66	100
3 FME standardisation	100	0	100
4 FCE level	25	75	100
5 CAE level	58.33	41.66	100
6 FME level	50	50	100
7 FCE academic	8.33	91.66	100
8 CAE academic	50	50	100
9 FME academic	81.81	18.18	100

Part 3 presented in Table 4.17 aimed at investigating whether the participants found the CEFR descriptors workable and relevant to their context. Whereas the participants

(91.66%) were satisfied with the "least able B2" definition they came up with as a group, the results of this section show that they were not convinced about the context-free nature of the descriptors in that they could also be used in academic contexts (41.66%). Moreover, for them the descriptors did not cater for all aspects of the skill of reading (75%). These issues are directly linked to context validity and cognitive validity. For instance, the skill of making inferences, which could be considered an academic skill, does not show itself consistently in the B2 descriptors of the illustrative scales for reading although CEFR is context-free, in that it can be used in all contexts including the academic context.

Table 4.17 Proportions of Responses in Two Categories – CEFR Descriptors

Question	Strongly Agree / Agree	Disagree / Strongly Disagree	%
1 LAB2	91.66	8.33	100
2 academic	58.33	41.66	100
3 context free	50	50	100
4 aspects of reading	25	75	100

Finally, the purpose of the last part of the questionnaire was to evaluate the contributions of the Reading paper standardisation to the COPE examination. The findings, presented in Table 4.18, were quite similar to those of the writing paper. Everyone agreed that the process forced them to reconsider the level of COPE and helped them understand the level of the exam better. It also pointed to areas for revision, resulting from features of levels specified in the CEFR descriptors, if the level of the exam were to be increased (80%). In addition, it was felt that the standard setting had a positive impact on the reliability (90%) and the validity (100%) of the COPE examination.

Table 4.18 Proportions of Responses in Two Categories – CEFR Linking Process

Question	Strongly Agree / Agree	Disagree / Strongly Disagree	%
1 reconsidering level	100	0	100
2 understanding level	100	0	100
3 increasing standards	80	20	100
4 reliability	90	10	100
5 validity	100	0	100

*1 person did not answer this section (N=10).

4.4.4.2 Field Notes

(a) Writing Standardisation

The coding scheme developed at the familiarisation stage was also employed for the field notes at the standardisation stage and the frequency of the codes are presented in Table 4.19. The table suggests that mainly four aspects of validity play a significant role in the CEFR writing standardisation: criteria with a frequency of 341, understanding the level of the exam (282), task demands (207) and language knowledge (164) respectively in order of emphasis. These aspects reflect the very nature of CEFR standardisation that it involves understanding the level of a writing task by considering what it requires test takers to do (task demands) and what language knowledge is needed to fulfil these demands through the use of the CEFR scales (criteria).

As the analysis of the field notes shows, the main emphasis of the writing standardisation was on criteria, a parameter of scoring validity; and task demands and language competence, parameters of cognitive and context validity. It could be argued, therefore, that the writing standardisation involved the relationship between context, cognitive and scoring validity, which is the core of construct validity as argued by Weir (2005a).

Table 4.19 Frequencies – Writing Standardisation

Themes	Descriptions	Frequencies
Test taker	Experiential	3
Context validity	Task design	66
	Task demands	207
Cognitive validity	Language knowledge	164
	Content knowledge	1
Scoring validity	Criteria	341
	Prompt	17
Implications on the level of the exam	Understanding of level	282

(b) Reading Standardisation

The analysis of the reading standardisation field notes are summarised in Table 4.20, which suggests that throughout the standardisation of reading, the participants were engaged in forming links between the criteria (251) – the CEFR scales – and the reading tasks (106), with some emphasis on the language knowledge required to complete the tasks. To be more explicit, task design features such as task type, purpose or distractors and task demands such as the text length and the linguistic requirements of a task formed the core of judgments made while relating the reading tasks or items to the criteria.

Table 4.20 Frequencies – Reading Standardisation

Themes	Descriptions	Frequencies
Context validity	Task design	97
	Task demands	106
Cognitive validity	Language knowledge	61
	Content knowledge	9
Scoring validity	Criteria	251
	Items	23
Implications on the level of the exam	Intended level	3
	Understanding of level	46
	Increasing the standards	2

Therefore, the field notes suggest that the reading standardisation also highlights the relationship between context, cognitive and scoring aspects of validity.

4.4.4.3 Statistical Analyses

(a) Writing Training

During the training part of the writing standardisation, the judges were asked to use Table 5.8, Written Assessment Criteria Grid, from the Manual (Council of Europe, 2003: 82) in order to assign levels to the samples provided. The internal consistency of judgments and the level of agreement in the writing training were high with an alpha index of .84 and an ICC of .80 as shown in Table 4.21.

Table 4.21 Agreement and Consistency of Judges – Writing Training

Writing	Inter-rater reliability			Alpha	ICC*
	Mean	Min	Max		
Training	6.5769	5.7500	7.5000	.8499	.8015

A comparison of the actual levels of the written samples to the levels assigned by the judges can be made by looking at Figure 4.2. The numbers under the NSample column on the right indicate the training samples used. The values under the Observed Average column indicate the average levels assigned by the judges out of 10 for writing. Both the Observed Average and the logit scale values given under the Measure column (column 5) show that the CAE exemplar with an average of 7.2 and a logit scale value of .93 has the highest level of difficulty and the first FCE exemplar, average 5.2 and logit scale value of -.54, has the lowest level of difficulty. However, even though the judges were able to correctly differentiate the C1 level paper from B2 level papers, there was an issue with one of the B2 level papers, that is, the judges assigned a lower level (B1) to this B2 level sample paper. The logit scale value of FCE1 is -.54 whereas those

of FCE2 and IELTS are .75. As was discussed earlier under 4.4.4.1, the FCE samples were found to be unsuitable for an academic context. The participants might have found the FCE1 sample, a short letter, easier than the others that were all longer texts. According to the Manual (2003: 75), "the vast majority of participants should agree on the level, with the spread not exceeding one and a half levels", in which case the judgments made for the FCE1 exemplar did not pose any problem and the judges were highly standard. An average of 5.2 corresponds to the B1 level and the judges were only 1 level off. However, these limits suggested by the Manual were considered too broad for reliability purposes and thus not tolerated because a one and a half band change in level is too much for setting cut scores, especially in the BUSEL context, where the goal is to set a cut score at one level (B2) only.

Figure 4.2 Sample Measurement Report – Writing Training

Obsvd Score	Obsvd Count	Obsvd Average	Fair-M Avrage	Model Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	NSample
68	13	5.2	5.21	-.54	.34	.90	.0	.95	.1	.95	FCE1 B2
94	13	7.2	7.24	.93	.21	1.15	.5	1.10	.3	.87	CAE C1
90	13	6.9	6.91	.75	.21	1.14	.5	1.11	.4	.96	FCE2 B2
90	13	6.9	6.91	.75	.21	.71	-.9	.70	-.9	1.31	IELTS B2
85.5	13.0	6.6	6.57	.47	.25	.98	.0	.96	.0		Mean(Count4)
10.2	.0	.8	.79	.59	.05	.18	.6	.17	.6		S.D.(Populn)
11.8	.0	.9	.92	.68	.06	.21	.7	.19	.7		S.D.(Sample)

Model, Populn: RMSE .25 Adj (True) S.D. .53 Separation 2.13 Reliability .82
Model, Sample: RMSE .25 Adj (True) S.D. .63 Separation 2.53 Reliability .86
Model, Fixed (all same) chi-square: 14.5 d.f.: 3 significance (probability): .00
Model, Random (normal) chi-square: 2.5 d.f.: 2 significance (probability): .28

Figure 4.3, which illustrates how the judges performed in the training, reveals high reliability of the judgment process with a value of .00 as indicated in the first row of the bottom part in the figure labelled Reliability (not inter-rater). As briefly explained earlier in section 4.2.4.3 of this chapter, the Rasch reliability index is

> "a rather misleading term as it is *not* an indication of the extent of agreement between raters (the traditional meaning of reliability indices between raters) but the extent to which they really differ in their level of severity. High reliability indices in this table indicate real differences between raters, not in their overall ranking of candidates, but in the actual levels of scores assigned to them." (McNamara, 1996: 140)

Therefore, the reliability index needs to be low for raters, that is, judges in our context (Linacre, 2007: 149). The reliability of .00 suggests that there is no difference in rater severity. This is also supported by the fact that the chi square of 7.8 with 11 degrees of freedom (d.f) is not statistically significant (p= .92). The Chi-square in rater analysis tests the hypothesis "Can these raters be thought of as equally lenient?" In other words, "Is there a statistically significant rater effect?" (Linacre, 2007). In this case, there is no significant difference among the judges. However, when the fit statistics are analysed, judges (denoted as R in Figure 4.3) 1,3,4,5,9 and 11 are considered as misfitting. An infit between 0.4 and 1.2 is considered reasonable by Linacre and Wright (1994). Considering the total number of judges in the training, this is a very high figure (6 out of 13). However, the infit values of 3 of these judges (1,9 & 11) are below 0.4, which might be due to the nature of the samples used for training. That is, 3 samples out of 4 were at B2 level and this might have led the judges to overuse the middle category (B2). Then the infit values would look both combined together with a relatively small number of observations and a limited range of scores, even one observation away will push the infit down. In addition, the expectation is that the judges will agree. Therefore, a low infit is unlikely to be an issue. If this were a big test with big numbers, this would be a problem. In figure 4.1, the sample measurement report, there is 1 paper assigned at B1

level, 2 papers at B2 and another one at C1 level. The other 3 judges (3, 4 & 5) are inconsistent in their judgements. The performance of these judges was monitored in the standard setting session and the results will be discussed in the following section.

Figure 4.3 Rater Measurement Report – Writing Training

Obsvd Score	Obsvd Count	Obsvd Average	Fair-M Avrage	Model Meas	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Exact Agree. Obs %	Exact Agree. Exp %	NuR
26	4	6.5	6.31	.06	.41	2.62	2.0	3.10	2.1	.12	14.6	38.3	3
25	4	6.3	6.05	.23	.42	2.20	1.6	2.09	1.3	-.11	12.5	38.0	4
24	4	6.0	6.05	.23	.42	1.50	.8	1.72	1.0	.44	20.8	38.0	7
26	4	6.5	6.31	.06	.41	1.58	1.0	1.32	.6	1.08	35.4	38.3	5
26	4	6.5	6.31	.06	.41	1.14	.4	.96	.1	1.35	37.5	38.3	8
30	4	7.5	7.51	-.61	.43	.72	-.2	.73	-.2	1.19	29.2	33.4	12
30	4	7.5	7.51	-.61	.43	.62	-.4	.62	-.4	1.53	37.5	33.4	13
25	4	6.3	6.05	.23	.42	.68	-.3	.57	-.3	.53	31.2	38.0	6
28	4	7.0	6.89	-.27	.41	.57	-.6	.52	-.6	1.56	39.6	37.1	2
28	4	7.0	6.89	-.27	.41	.57	-.6	.52	-.6	1.56	39.6	37.1	10
23	4	5.8	5.56	.61	.47	.34	-1.0	.27	-.8	.97	29.2	35.6	1
25	4	6.3	6.05	.23	.42	.10	-2.3	.09	-1.8	1.10	35.4	38.0	9
26	4	6.5	6.31	.06	.41	.02	-3.2	.04	-2.4	1.83	45.8	38.3	11
26.3	4.0	6.6	6.45	.00	.42	.97	-.2	.96	-.2		Mean(Count:13)		
2.1	.0	.5	.57	.34	.02	.76	1.4	.85	1.2		S.D. (Populn)		
2.1	.0	.5	.59	.35	.02	.79	1.5	.88	1.3		S.D. (Sample)		

Model, Populn:RMSE .42 Adj(True) S.D. .00 Separation.00 Reliabil(not inter-rater) .00
Model, Sample:RMSE .42 Adj (True) S.D. .00 Separation .00 Reliabil (not inter-rater) .00
Model, Fixed (all same) chi-square: 7.8 d.f.: 12 significance (probability): .80
Model, Random (normal) chi-square: 5.2 d.f.: 11 significance (probability): .92
Rater agreement opportunities: 312 Exact agreements: 98 = 31.4% Expected: 115.6 = 37.0%

(b) Writing Standard Setting

The Standard Setting tasks are similar to the Training tasks and the same analyses were employed at this stage. All three sessions will be discussed separately.

Session 1

Table 4.22 demonstrates that the consistency of judgements and the level of agreement in the writing standard setting were very high and that the values improved after each round.

Table 4.22 Agreement and Consistency of Judges – Writing Standard Setting Session 1

Writing Stand Sett Session 1	Inter-rater reliability			Alpha	ICC*
	Mean	Min	Max		
Round 1	7.0923	5.8000	7.6000	.9692	.9704
Round 2	7.0000	6.6000	7.4000	.9891	.9885
Round 3	6.9231	6.8000	7.2000	.9963	.9950

*Statistically significant at level p≤.01

Table 4.23 shows the actual COPE scores of the samples and their CEFR levels at the end of the standard setting. In the COPE writing criteria, which gives a holistic grade out of 5, a pass grade is considered to be a 3. The results of the standard setting indicate that bands 3 and 4 on the COPE criteria are within the B2 band and that the 5 band is equivalent to B2+.

Table 4.23 COPE Scores and Their CEFR Levels

SAMPLE	COPE SCORE	CEFR LEVEL
1	2	B1
2	3	B2
3	4	B2
4	5	B2+
5	3	B2

Two different types of Rasch analyses were carried out on the data set. The first, using the model ?,?,R6, estimates the probability of a rating by any rater on a scale of 6. The second, using the hybrid model line ?,#,R6, simply highlights the rater facet by producing separate output tables for each variable in the facet marked #, in this case the rater. The difference between these two analyses was that the former investigates the performance of raters when they are modeled "to share a common understanding of the rating scale" and the latter when they are modeled to "have a personal understanding of

the rating scale (Linacre, 2007: 215). The results that follow are from the first type of analysis. The hybrid model was also run to see what the severity differences were between the judges, but, since they were so close to each other in terms of severity, the analysis was rejected by FACETS. This means that the judges were not behaving as independent raters but as sharing a common understanding of the CEFR scale for writing and thus there is rater agreement. The first model was used for all the other standard setting sessions.

Table 4.24 shows the performance of the judges over three rounds. The range of logits in round 1, round 2 and round 3 was 4.99, 5.44 and 8.49 repectively. This shows that after each round the judges drifted from each other on the logit scale rather than getting closer. However, when the number of samples analysed and the fact that the judges were intended to "act like machines" are considered, slight differences in judgements are reflected as high values in the Rasch analysis. By "acting like machines", Linacre (2007) wants to emphasize that the raters are all marking in the same way. This can clearly be seen in Table 4.25 where the descriptive statistics are presented. It should also be clarified here that the argument of a slight difference in judgments being shown as high values is equally valid for the fit statistics. Therefore, judges 2 and 7 with infit values of 1.95 cannot be considered inconsistent.

Table 4.24 Rater Measurement Report – Writing Standard Setting Session 1

	Round 1			Round 2			Round 3		
Rater	Logit	SE	Infit MnSq	Logit	SE	Infit MnSq	Logit	SE	Infit MnSq
1	-.38	.74	1.02	-1.87	1.13	1.45	2.63	4.44	.02
2	1.52	.81	.30	1.64	1.33	.00	-1.66	1.49	1.95
3	.87	.82	.04	-.22	1.47	.22	-1.66	1.49	.51
4	.22	.80	.33	-.22	1.47	.22	2.63	4.44	.02
5	3.69	.76	2.46	3.57	1.59	.21	2.63	4.44	.02
6	-.38	.74	.14	-1.87	1.13	.19	-5.86	4.37	.02
7	-.87	.68	.20	-.22	1.47	.22	-1.66	1.49	1.95
8	-.38	.74	.14	-1.87	1.13	.19	-1.66	1.49	.51
9	-.38	.74	.14	-.22	1.47	2.65	-1.66	1.49	.51
10	-.87	.68	2.30	-.22	1.47	.22	2.63	4.44	.02
11	-.87	.68	.55	-.22	1.47	.22	2.63	4.44	.02
12	-.87	.68	.55	3.57	1.59	.21	2.63	4.44	.02
13	-1.30	.64	2.33	-1.87	1.13	.19	-1.66	1.49	.51
Mean	.00	.73	.81	.00	1.38	.48	.00	3.07	.47
SD	1.30	.06	.89	1.81	.17	.71	2.66	1.46	.67

As illustrated in Table 4.25, the range of levels assigned to the samples is 0 or 1, which is very low meaning that the judges were mostly in agreement regarding the levels assigned to the samples.

Table 4.25 Descriptive Statistics from Writing Standard Setting Session 1

Writing	N	Mean	Median	Mode	SD	Range	Min	Max
1	11	5	5	5	0	0	5	5
2	11	7	7	7	0	0	7	7
3	11	7.35	7	7	0.49	1	7	8
4	11	8.21	8	8	0.42	1	8	9
5	11	7	7	7	0	0	7	7

This is also supported by the reliability values in Table 4.26, which reveals that the reliability of the judgement process increased after each round. In addition, chi-square values showed significant difference in rounds 1 and 2 and no significant (.91) difference among the judges at the end of round 3.

Table 4.26 Rater Consistency – Writing Standard Setting Session 1

	Reliability	Chi-Square	d.f	Significance
Round 1	.68	39.8	12	.00
Round 2	.41	21.6	12	.04
Round 3	.00	6.1	12	.91

The judgement process can also be investigated by looking at the rater agreement statistics in Table 4.27, which shows that the observed agreements in each round are higher than the expected agreements. If the percentage of observed agreements is higher than the percentage of expected agreements, this means that the judges may be behaving like 'rating' machines (Linacre, 2007: 150), which is desirable in such benchmarking studies.

Table 4.27 Rater Agreement Opportunities – Writing Standard Setting Session 1

	Round 1	Round 2	Round 3
Observed Agreements	215=55.1%	302=77.4%	242=77.6%
Expected Agreements	205.2=52.6%	294.3=75.5%	236.3=75.7%

Rater Agreement Opportunities: 390

The slight change in observed agreements from 77.4 in Round 2 to 77.6 in Round 2 demonstrates that the judgment process became stable even after Round 1. The problems in terms of agreement were fixed after Round 1. The implication of this could be that the first two rounds were the most significant rounds in the standard setting and that the third one is carried out for confirmation purposes.

Session 2

Data from session 2 were analysed using MFR, which showed that there are overlaps between the levels, which is acceptable and inevitable. This is also the case with the CEFR levels in that "one should be careful about interpreting sets of levels and scales of

language proficiency as if they were a linear measurement scale like a ruler. No existing scale or set of levels can claim to be linear in this way" (Council of Europe, 2001: 17). Even though the levels seem to be equidistant on the CEFR scales, each level is situated halfway to the following level (ibid.)

However, the data are not presented here as the data set included several anomalies. For instance, COPE band 5 ranging from 24 to 27 points included papers that were assigned LAB2, or band 4 had papers labelled as low as B1+. Considering the low rater reliability of .68 and that 4 out of 11 judges were inconsistent in their judgments, it was decided to hold another session, as also mentioned in section 4.4.2.1. The low reliability might be due to the time between two sessions, in that session 2 was held two years after the first one, as explained again in section 4.4.2.1. Participants might have needed further training on the CEFR writing scales to be able to successfully complete session 2.

Session 3

Table 4.28 summarises the analyses carried out for the two rounds of judgments made during session 2. The consistency of the judges and the levels of agreement in session 3 considerably improved in the second round but were lower than those of session 1. However, it should be noted that for a sample size of 11 judges and 10 samples, the alpha and ICC reported in Table 4.28 are considered high because with such small sample sizes especially when the sample under study is not heterogeneous, (Frisbie, 1988; Traub & Rowley, 1991), the reliability decreases as the score variance becomes smaller. In this case, low reliability might mean that the group had assigned similar levels to the samples. This, in fact, raises questions regarding the use of classical

reliability theories suggested by the Manual in standard setting situations where high consensus is sought. As mentioned earlier in section 3.7.1.1, Kaftandjieva (2004) argues that correlational analyses are not appropriate for standard setting purposes as a perfect correlation can be achieved with no rater agreement. Alpha and ICC are also types of correlations and besides the shortcoming mentioned by Kaftandjieva, such analyses also poses problems to users at interpretation level. Correlation indexes are affected by sample size as well as variance, which is an indication of levels of agreement. Low variance, showing a high level of agreement, tends to result in a low reliability figure.

Table 4.28 Agreement and Consistency of Judges – Writing Standard Setting Session 3

Writing Stan Sett Session 3	Inter-rater reliability			Alpha	ICC*
	Mean	Min	Max		
Round 1	2.9545	1.3636	3.9091	.0308	.0871
Round 2	3.0000	1.4545	4.7273	.6929	.6454

*Statistically significant at level $p \leq .01$

Table 4.29 presents the actual COPE scores of the samples and their CEFR levels. It reflects the results of both the 12 scripts assessed as a group and 8 done individually, statistical analysis of which is presented here. Whereas a confident claim can be made regarding the B2 level at a score of 21 on the COPE writing paper, there is a grey area between 18 – 20 points where some papers falling into this range were categorized as a least able B2 performance. Therefore, the writing cut score has been set at 21 as papers with a minimum score of 21 are consistently categorized as B2 level performance.

Table 4.29 COPE Grades and Their CEFR Equivalents

COPE BAND	SCORE RANGE	CEFR LEVEL
6	28-30	
5	24-27	
4	21-23	LAB2/B2
	18-20	B1+/LAB2
3	12-17	B1+
2	6-11	
1	1-5	
0	0	

Table 4.30 demonstrates the performance of judges over two rounds in session 3. Similar to the findings in session 1, the range of logits in round 2 (4.16) is higher than round 1 (1.59) due to the reason explained above under session 1.

Table 4.30 Rater Measurement Report – Writing Standard Setting Session 3

	Round 1			Round 2		
Rater	Logit	SE	Infit MnSq	Logit	SE	Infit MnSq
1	-.91	.52	1.02	-1.43	.73	1.21
2	.41	.52	.69	.46	.66	.74
3	-.65	.52	.70	-.92	.72	.23
4	.68	.52	2.86	2.73	.75	.92
5	-.91	.52	.32	-.44	.69	.63
6	.41	.52	.67	.46	.66	.91
7	.14	.52	.64	.02	.67	.80
8	.14	.52	.56	.46	.66	.75
9	.68	.52	1.10	-.44	.69	1.45
10	.14	.52	1.42	-.92	.72	.23
11	-.12	.52	.83	.02	.67	.88
Mean	.00	.52	.98	.00	.69	.80
SD	.56	.00	.66	1.05	.03	.34

Unlike session 1, however, for three samples (samples 4, 6, and 8), the range of levels assigned is 2 indicating one CEFR level of difference in judgments because the levels used during the process also included the plus (+) levels. For instance, in this case the range being 2 means a change in level between B1 and B2 because of the B1+ level in between. In cases where the range is 2, it might be safer to look at the median and the

mode and not just the mean to decide on the level of a paper. Whereas for samples 6 and 8 the decision is an easier call, for sample 4 it is more complicated due to the differences in the median and mode values. In this case, a more conservative approach was followed and the sample was assigned a level of 3 (B2).

Table 4.31 Descriptive Statistics from Writing Standard Setting Session 3

Writing	N	Mean	Median	Mode	SD	Range	Min	Max
1	11	3	3	3	0	0	3	3
2	11	3.90	4	4	0.30	1	3	4
3	11	1.45	1	1	0.52	1	1	2
4	11	3.63	4	3	0.67	2	3	5
5	11	2.72	3	3	0.46	1	2	3
6	11	3.72	4	4	0.64	2	3	5
7	11	4.72	5	5	0.46	1	4	5
8	11	2.27	2	2	0.64	2	1	3
9	11	2.27	3	3	0.46	1	2	3
10	11	1.81	2	2	0.4	1	1	2

When the MFR reliability values in Table 4.32 are analysed, although the reliability in round 2 is lower, the chi-square analysis indicated no significant difference among the judges in both rounds.

Table 4.32 Rater Consistency – Writing Standard Setting Session 3

	Reliability	Chi-Square	d.f	Significance
Round 1	.14	12.7	10	.24
Round 2	.57	22.8	9	.53

Table 4.33 supports the finding that the judges were acting similarly. The observed agreements in both rounds are higher than the expected agreements. As explained for session 1, this means that the judges were acting like ‘rating’ machines.

Table 4.33 Rater Agreement Opportunities – Writing Standard Setting Session 3

	Round 1	Round 2
Observed Agreements	237=43.1%	317=57.6%
Expected Agreements	240.7=43.8%	299.9=54.5%

Rater Agreement Opportunities: 550

(c) Reading Training

The data on training comes from the first and the second reading standardisation sessions. In the first session, the internal consistency of judgments and the level of agreement in the reading training were very high as shown in Table 4.34. However, as explained in section 4.4.2.2, this might be due to the fact that the judges were able to identify the examinations the samples came from and therefore, might have made their judgments based on this knowledge.

Table 4.34 Agreement and Consistency of Judges – Reading Training Session 1

Reading	Statistics			Alpha	ICC*
	Mean	Min	Max		
Training	.3269	.2500	.7500	.9791	.9773

*Statistically significant at level p≤01

In the second session, the statistical analysis of the judgments was troublesome for two possible reasons. The first reason is that two out of four items showed no variance, which caused the reliability to be distorted as seen in Table 4.35. No variance meant that all the judges assigned the same levels to these items. The second reason could be the influence of the small sample size on the reliability estimate.

Table 4.35 Agreement and Consistency of Judges – Reading Training Session 2

Reading	Statistics			Alpha	ICC*
	Mean	Min	Max		
Training Yes/No	0,833	0,833	0,833	.2000	.2000

*Statistically significant at level p≤01

The many-facet Rasch analysis, revealing a more reliable estimate of reliability, presented in Figure 4.4 showed that the judgment process in the training was highly reliable with a value of .00.

Figure 4.4 Rater Measurement Report – Reading Training Session 2

Total Score	Total Count	Obsvd Average	Fair-M Avrage	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Correlation PtMea	Correlation PtExp	Exact Agree. Obs %	Exact Agree. Exp %	Nu Rater
2	4	.5	.50	.00	1.41	1.00	.0	1.00	.0	1.00	.71	.49	.0	50.0	5 5
2	4	.5	.50	.00	1.41	1.00	.0	1.00	.0	1.00	.71	.49	.0	50.0	8 8
1	4	.0	.15	(1.73	1.98)	Maximum					.00	.00	.0	.0	1 1
1	4	.0	.15	(1.73	1.98)	Maximum					.00	.00	.0	.0	2 2
1	4	.0	.15	(1.73	1.98)	Maximum					.00	.00	.0	.0	3 3
1	4	.0	.15	(1.73	1.98)	Maximum					.00	.00	.0	.0	4 4
1	4	.0	.15	(1.73	1.98)	Maximum					.00	.00	.0	.0	6 6
1	4	.0	.15	(1.73	1.98)	Maximum					.00	.00	.0	.0	7 7
1	4	.0	.15	(1.73	1.98)	Maximum					.00	.00	.0	.0	9 9
1	4	.0	.15	(1.73	1.98)	Maximum					.00	.00	.0	.0	10 10
1	4	.0	.15	(1.73	1.98)	Maximum					.00	.00	.0	.0	11 11
1	4	.0	.15	(1.73	1.98)	Maximum					.00	.00	.0	.0	12 12
1.2	4.0	.1	.21	1.44	1.88	1.00	.0	1.00	.0		.12				Mean (Count: 12)
.4	.0	.2	.13	.64	.21	.00	.0	.00	.0		.26				S.D. (Population)
.4	.0	.2	.14	.67	.22	.00	.0	.00	.0		.28				S.D. (Sample)

With extremes, Model, Popuin: RMSE 1.89 Adj (True) S.D. .00 Separation .00 Reliability (not inter-rater) .00
With extremes, Model, Sample: RMSE 1.89 Adj (True) S.D. .00 Separation .00 Reliability (not inter-rater) .00
Without extremes, Model, Popuin: RMSE 1.41 Adj (True) S.D. .00 Separation .00 Reliability (not inter-rater) 1.00
Without extremes, Model, Sample: RMSE 1.41 Adj (True) S.D. .00 Separation .00 Reliability (not inter-rater) .50
With extremes, Model, Fixed (all same) chi-square: 2.2 d.f.: 11 significance (probability): 1.00
Inter-Rater agreement opportunities: 2 Exact agreements: 0 = .0% Expected: 1.0 = 50.0%

(d) Reading Standard Setting

Session 1

Regarding reading session 1 only one set of data is presented to explain the problem experienced in this session, mentioned in section 4.4.2.2.

Table 4.36 summarizes the results of the reading standard setting based on the Yes/No method, which clearly reflects the problems experienced with the standard setting

methods. The mean for the reading paper was 0.77 which corresponds to a cut score of 27 out of 35. However, the standard deviation was very high and this was also reflected in the minimum (20) and the maximum (34) cut scores set by individual judges. Whereas one of the judge's cut score was as low as 20, another judge's cut score was as high as 34, with 14 points difference. Even though such problems are expected in standard setting sessions based on judgments, the discussion that followed the first round of judgments revealed that the judges were not clear about the least able B2 candidate definition and how to work with it. It became apparent that they were making judgments about the items based on completely different standards of their own. Consequently, it was agreed that a further standard setting session was required.

Table 4.36 Results of the Reading Standard Setting Session 1 – Yes/No Method

No. of items	Mean	St Dev	Cut score		
			Min	Max	Mean
35	0.77	0.40	20	34	27

Immediately after the session, some participants had an informal discussion to explore the problems experienced in the standard setting. Being biased by the exit level requirements of the institution, giving too much emphasis on test-taking strategies, considering the B2 level candidate rather than the least able B2 candidate were some of the problems identified through the discussion. This informal discussion was not part of the research but is worth mentioning here as it gave further insight into the findings of the statistical analysis of the reading standardisation.

Session 2

Table 4.37 presents the cut scores established as a result of two standard-setting methods. The mean is the cut score and the standard deviation shows the variability in

judges' mean ratings. The mean ratings for the two rounds of judgments in both methods were similar in variability, with the round 1 ratings slightly less variable than the round 2 ratings. When the two methods are compared, the mean ratings of the modified Angoff methods were slightly less variable than the Yes/No method ratings.

Table 4.37 Results of the Reading Standard Setting

		N	Mean	Median	Mode	SD	Range	Min	Max
Yes/No Method	R1	10	22.3	22	22	2.31	7	19	26
	R2	10	20.5	20.5	20	2.91	7	17	24
Modified Angoff	R1	10	19.61	19.6	19.5	1.74	6.1	15.8	21.9
	R2	10	19.09	19.24	19.4	2.08	7.7	14.3	22

At the end of round 2, based on the Yes/No method the judges decided that the least able B2 candidate should be required to attain a proportion correct of .61, which equates to a cut score of 21 in this 35-item reading test. Based on the modified Angoff method, the least able B2 candidate should be required to attain a proportion correct of 54.54, which equals to a cut score of 20 out of 35. The rounding of the scores is done according to the decisions outlined in section 4.4.2.2. The data are presented in Table 4.38, which also shows the adjusted cut scores based on the standard error of the mean.

Table 4.38 Reading Cut Score - Rounded and Adjusted

		N	Proportion	Cut score	Rounded	SE	Adjusted
Yes/No Method	R1	10	0.64	22.3	23	0.73	24
	R2	10	0.61	20.5	21	0.29	22
Modified Angoff	R1	10	56.06	19.61	20	0.55	21
	R2	10	54.54	19.09	20	0.65	22

It is interesting that the difference between the rounded and adjusted cut scores established at the end of round 1 and round 2 is high in the Yes/No method whereas they are the same in the modified Angoff method. When the two different standard-

setting methods are considered, the difference between the final cut score suggested is 1 point or in other words, one item.

As presented in Table 4.39, the internal consistency of judgments and the level of agreement are higher in the modified Angoff method.

Table 4.39 Agreement and Consistency of Judges – Reading Standard Setting

Reading		Descriptive Statistics			Alpha	ICC*
		Mean	Min	Max		
Yes/No Method	R1	.6371	.5429	.5429	.7283	.7271
	R2	.6086	.4857	.7714	.7901	.7920
Modified Angoff	R1	.56.05	45.14	62.57	.8754	.8737
	R2	54.54	40.85	62.85	.9223	.9217

*Statistically significant at level p≤01

Table 4.40 reports the Rasch analyses of the judgment process for both standard setting methods. The narrow range of logits (1.24, 1.95, 1.00, 1.60) shows that the judges were relatively close to each other on the logit scale, which means that there was not much difference among them in terms of leniency. When the infit indices are analysed, rater 9 in the Yes/No method, rater 2 for round 1 and rater 8 for round 2 of the modified Angoff method seem to be misfitting. For standard setting purposes, even a single misfitting judge is not desirable. However, when the raw data is analysed, it can be seen that these judges demonstrated only slight drifts that were reflected as big changes in the many-facet Rasch analysis.

Table 4.40 Rasch judge consistency – Reading Standard Setting

	Yes/No Method						Modified Angoff					
	Round 1			Round 2			Round 1			Round 2		
Judge	Logit	SE	Infit MnSq	Logit	SE	Infit MnSq	Logit	SE	Infit MnSq	Logit	SE	Infit MnSq
1	.07	.42	1.06	.07	.43	1.04	.01	.13	1.00	-.05	.15	1.19
2	.56	.41	1.19	.80	.43	1.18	.18	.13	1.31	.38	.14	.88
3	-.68	.46	1.12	-1.15	.49	1.20	.00	.13	.94	.03	.15	.89
4	-.48	.44	1.07	-.50	.45	1.18	-.40	.14	.50	-.29	.15	.57
5	.24	.41	1.12	.25	.43	.79	.03	.13	.74	.07	.15	.67
6	.07	.42	.77	.07	.43	.92	-.16	.13	.80	-.14	.15	.81
7	.24	.41	.88	.44	.43	.90	.11	.13	.72	-.27	.15	.69
8	.07	.42	1.02	-.50	.45	.84	.18	.13	1.46	-.05	.15	1.51
9	.40	.41	1.34	.25	.43	1.44	-.34	.14	1.09	-.64	.15	1.21
10	-.48	.44	.80	.25	.43	.67	.60	.12	.90	.96	.14	.91
Mean	.00	.42	1.01	.00	.44	1.02	.00	.13	.95	.00	.15	.94
SD	.39	.02	.18	.53	.02	.22	.27	.00	.27	.41	.00	.28

4.4.5 Standardisation stage summary of research findings

This stage of the CEFR linking process evidently reflects what standard setting is all about and is defined as "a decision making process aiming to classify the results of examinations in a limited number of successive levels of achievement (proficiency, mastery, competency) by Kaftandjieva (2004: 2). Kaftandjieva also emphasizes the distinction between *Content standards* that refer to the curriculum and *Performance standards* that refer to explicit definitions of what students must do to demonstrate proficiency at a specific level on the content standards. (CRESST Assessment Glossary, 1999). Standardisation or standard setting could also be described as the process of turning content standards into performance standards. In the case of the COPE CEFR linking project, the act of transforming the content standards as defined by the CEFR into performance standards for the COPE examination following the procedures suggested by the Manual was successful. This is supported by the judge consistency and agreement statistics as well as reliability indices.

From a research perspective, the analysis of the questionnaires showed that both writing and reading standardisation sesssions were effective, though the reading standardisation was perceived slightly less effective than writing. Effectiveness is an important indicative of the reliability and validity of standard setting, which contributes to the overall validity of an examination. This issue will be further discussed in the concluding chapter. The questionnaires also helped identify three main issues with the standard setting process. Firstly, in the writing standardisation, the FCE samples were regarded as unsuitable for an academic context. Secondly, the Finnish reading calibrated items were the only samples that were found to be useful as they reflected the COPE task type and also because the ESOL items were predictable in terms of level. Thirdly, the participants did not agree that the CEFR scales were context-free. These three issues highlight some lacks in the Manual. A better variety of samples should accompany the Manual to guide users in their linking process. In addition, guidance or examples of how the CEFR scales can be used in different contexts should be provided. In terms of the quality of the COPE examination, the participants indicated that both writing and reading standardisation sessions contributed to the validity and reliability of the examination.

The field notes for writing and reading both highlighted the relationship between context, cognitive and scoring aspects of validity. If data are collected systematically at the standardisation stage of the linking process, they can be used as evidence towards these aspects of validity. The data also suggest that the standardisation stage helped participants understand the level of the COPE examination better, particularly in the writing standardisation.

The standardisation process also turned out to be enlightening with respect to the use of statistics for both setting cut scores and analysing the quality of the standard setting process. First of all, it demonstrated that the use of multiple standard setting methods is valuable but challenging on the part of the judges. It is valuable because a cut score originated from one method can be confirmed by another. It is challenging as different methods might require different mindsets as was the case in reading standard setting session 1 where the judges were asked to use three different methods. Secondly, reliability analysis provides information as to which method is more reliable although it might change from one group of judges to another. Thirdly, the use of different statistical tools to analyse the judgments makes interpretation of the findings and taking decisions easier. Furthermore, the role of MFR in standard setting is invaluable as it allows for looking at standard setting from three angles. The first one is the samples. As MFR produced a fair average level for each sample and the associated error taking the judge effect into consideration, samples or items can be assigned CEFR levels in a more reliable manner. The second one is the judges. MFR analysis allows for investigating judge consistency on an individual level and agreement on a group level. This enables users to either consider judge differences or completely leave some of them out of the data set if necessary. The last one is the judgments in that the reliablity of the judgments is calculated in a more reliable way.

However, MFR analysis for standard setting purposes is not without flaws. As discussed earlier when presenting standard setting results for both writing and reading. Although Rasch analysis works well with small sample sizes (Lord, 1980), at times when the judges are in high agreement, the logit scale values of fit statistics might be affected.

Therefore, not only is the use of a variety of statistical tools to analyse standard setting data invaluable, but it is also vital to interpret the findings wisely.

As regards the COPE examination, not only did the standardisation stage confirm the level set at the specification stage but more importantly it confirmed the level initially set and traditionally sustained.

4.5 Empirical validation stage

4.5.1 The Manual approach

Empirical validation is crucial in order to provide evidence that the exam itself is valid and that the claims made at the end of the specification stage and the standardisation stage of the linking process are reliable and can be confirmed. The process outlined in the Manual sees empirical validation as encompassing two parts: internal validation and external validation (Council of Europe, 2003).

Internal validation is about establishing the quality of a test, which is also a pre-requisite for linking to the CEFR. The Manual defines features of a quality test as having good items, reflecting the intended level, providing reliable results, measuring what it claims to measure and having proper administration and marking systems (ibid). The Manual suggests that the following statistical analyses could be carried out for internal validation where relevant and appropriate:

- Classical Test Theory
- Qualitative Analysis Methods such as reflection, analysis of samples, analytical frameworks, and feedback methods
- Generalisability Theory

- Factor Analysis
- Item Response Theory

External validation, on the other hand, is essential in verifying the relationship of the cut scores set for the exam with the CEFR levels themselves (ibid). It mainly involves correlation; that is, correlating the scores on the exam in question with those of an acceptable measure of the intended concept, and matching classifications to the CEFR levels, which is converting a quantitative test score into a qualitative category represented by the CEFR levels (ibid).

4.5.2 Overview of the empirical validation stage of the project

Validation must be based on a validation framework but the Manual does not propose any such framework or recommend users to employ a theory. The approach followed by the COPE CEFR linking project was to provide a complete validation argument for the COPE examination following a model, a decision also taken for the City and Guilds CEFR alignment project (O'Sullivan, 2009a). The COPE CEFR linking project aimed to accumulate evidence to support the empirical validation argument both for the COPE exam and its link to the CEFR based on Weir's validation framework (2005a) for the reasons discussed in Chapter 2 section 2.4.2. The framework is comprised of the following aspects of validity:

- Test taker characteristics
- Context validity
- Cognitive validity
- Scoring validity
- Consequential validity

- Criterion-related validity

The framework was used to collect evidence for what are referred to as internal and external validity in the Manual (Council of Europe, 2003) and suggestions given by the Manual are also taken into consideration for internal validation, which is presented in the scoring aspect of validity in Weir's framework; and external validation, which is the criterion-related aspect of validity in Weir's validation model. In other words, Weir's validation framework encompasses aspects of validity covered through the Manual suggestions, with a focus on other aspects of validity such as context and consequential. Qualitative methods and factor analysis to investigate the thought processes of test takers and the dimensions of ability measured through an examination could not be undertaken in this study due to lack of resources and time constraints.

4.5.3 Empirical validation stage research procedures

The research procedures in this section include an interview with the project leader and statistical data. The interview was conducted to investigate the role of the empirical validation stage in a CEFR linking study and whether the stage contributed to the validity of the COPE examination. Statistical evidence was gathered as part of the project to put forward a validation argument for the COPE examination, which was also used as part of this research as it provides information as to whether the statistical tools in fact contributed to the validity argument of the examination. From a research perspective, the results of the statistical analyses provide information as to whether empirical validation, as suggested by the Manual, actually contributed to the validity of the COPE examination. The diversity of the methods used for empirical validation is restricted by certain institutional constraints which are explained in detail in 4.5.4.2.

4.5.4 Data analysis and findings concerning the empirical validation stage

4.5.4.1 Interview Data

The project leader was interviewed about the contributions of the empirical validation stage to the COPE examination. Three main areas were focused on in the interview: the role of the empirical validation stage in the linking process, the contribution of the stage to the validity of the COPE examination, and whether empirical validation helps maintain or adjust the level of the examination. The opinions of the project leader are presented here for each of the interview questions separately (See Appendix 3K for the interview coding scheme).

Firstly, in terms of the role of the empirical validation stage in the CEFR linking process, the project leader indicated that the empirical validation stage tells users where they can obtain statistical information about their examination and its link to some other criterion. She also stated that "if we think beyond that you're empirically validating your exam and the procedures throughout from the very beginning of the project. It is not the end stage it's something that's built in" (EV: 7-9). In other words, the whole linking process, not just the empirical validation stage, is validation of an examination.

Secondly, regarding the contribution of the empirical validation stage to the COPE examination in terms of different aspects of validity, the project leader indicated that the validation of the reading paper was done through comparing it with other Cambridge ESOL exams. Looking at the results of the calibration of FCE and COPE items, she further analysed the COPE reading paper and the FCE reading paper in terms of text level and language. Although the FCE texts seemed more challenging with more colloquial language, the COPE texts were more academic in nature and the level of tasks were similar. This was reassuring and "*it was a check that our understanding for*

COPE B2 level for reading was pretty much in line with the Cambridge B2 level" (EV: 35-37).

As for the writing paper, external or criterion-related validation was done through teacher judgments and it was an "eye-opener". Looking at the results of the comparison between teacher judgments and the COPE results,

> *"We've seen that a large portion of our students are not at B2 level and that has come out purely from the empirical validation stage and not all our teachers are ready to make those judgments. So it has helped us and particularly been valuable in looking at the level we expect and realistically seeing the number of students who get there so it's about teaching implications. We need to get more students up to that level. We wouldn't have found that out if we hadn't looked at the empirical validation stage. After this project we're quite confident about the standard we want our students to get to."* (EV: 49-56).

Thirdly, the last part of the interview looked at how the empirical validation stage would help to increase the level or maintain the standards of COPE. The project leader said that rather than increasing the level of the examination, the school needs to increase the standards students are getting to. Empirical validation stage helped in identifying texts or items that are below or at times above the expected level. Even though texts in the COPE item bank might have acceptable mean logit scale values, "*some of the items are too high and not necessarily at the level, we need to look at how we get the mean logit scale values.*"

The project leader seems to be emphasizing the need to further analyse texts and items in a test based on the results of the empirical validation stage to be able to set and maintain the level of an examination.

4.5.4.2 Statistical Analyses

Background information regarding the COPE examination, including its quality, was presented in section 3.2.2 of the Research Design chapter. It was indicated that COPE had gone through four revision cycles and the test specification were designed based on Weir's validation model. The quality of the examination is maintained through an IRT item banking system and monitoring marker performance is regular practice. In section 3.2.2, it was argued that these qualities of the COPE examination made it a legitimate examination for CEFR linking.

The validity argument for the COPE examination, as mentioned previously in the introduction of section 4.5.4, was based on Weir's validation model, which encompasses the Manual suggestions. In this section, the results of the empirical validation stage as stipulated by the Manual, i.e. internal validity (scoring) and external validity (criterion-related) are presented. Detailed information on other aspects of validity for the COPE examination can be found in Appendix 4F in Folder 7 of the accompanying CD.

The internal validation of the reading anchor test was investigated using classical item analysis and MFR was used for the internal validation of the writing paper. In terms of external validation, teacher judgments were used to develop decision tables for reading and writing; item facility values of the reading anchor test were correlated with the estimates of the project members who took part in the standardisation stage; and the reading anchor test was calibrated with FCE and CAE reading items.

a) Internal validation - Scoring validity of the COPE anchor test

The set of COPE sections that was used for CEFR linking purposes is referred to as 'the anchor test' from here onwards. Table 4.41 presents the descriptive statistics for the reading anchor test from the live running of the exam. It can be seen that the mean for the reading paper is slightly below the cut score established at the standard setting, which is 21. The reading paper has acceptable difficulty and point biserial values with moderate reliability. The population for the June administration is truncated as the candidates tend to be from the BUSEL programme. This would account for the relatively low reliability values reported here (the average reliability values for COPE over the past four years -2005 to 2009 - are 0.79 for reading). See accompanying CD folder 5 for ITEMAN, QUEST and FACETS outputs of the anchor test analyses.

Table 4.41 Scoring Validity of the Anchor Reading Test

N: 948	READING PAPER (out of 35)
Mean	18.652
Variance	20.721
St. Dev.	4.552
Skew	-0.061
Kurtosis	0.155
Min	3
Max	32
Median	19
Alpha	0.667
SEM	2.629
Mean P	0.533
Mean Item-tot	0.283
Mean Biserial	0.375

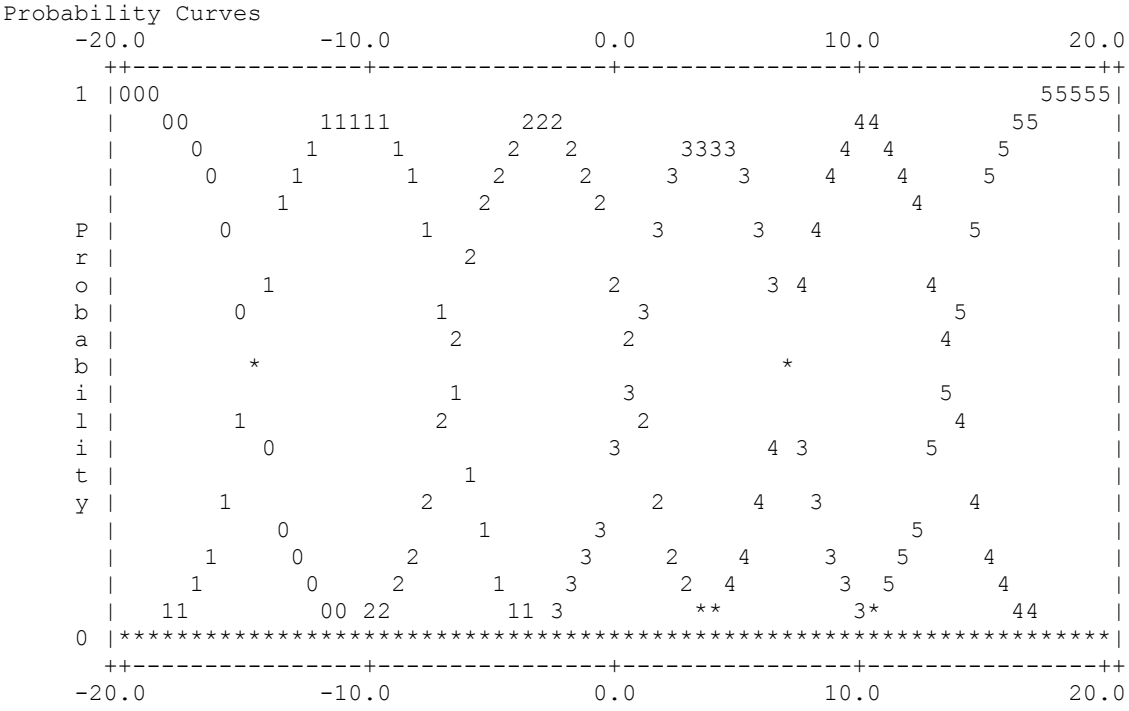
Table 4.42 presents evidence regarding the scoring validity of the anchor test writing paper. With a moderate reliability figure of .31, the rating of the anchor test was problematic due to the high number of inconsistent raters. The agreement statistics also supports the view that the marking was flawed, as observed agreements were less than

expected which indicates that the raters were not standard in their marking. The reason for this could be explained by the change in the marking system. Even though the institution has a core group of markers regularly trained, for this particular administration most of this group could not be involved in marking due to other work-related responsibilities. Thus, the marking was carried out by inexperienced markers for the most part. The COPE marking criteria, on the other hand, works well as the raters were able to easily differentiate between the bands in the criteria (Figure 4.5). The analysis of the writing paper reveals that quite intensive marker training is required.

Table 4.42 Scoring Validity of the Writing Paper

Rating reliability	.31 (MFR based index needs to be close to .00 indicating no real differences between rater)
Rater consistency	15 out of 23 raters were found to be inconsistent (based on MFR infit mean square statistics acceptable range 0.4-1.2
Observed agreements (out of 334 opportunities)	237 (71%)
Expected agreements (out of 334 opportunities)	246.1 (73.7%)

Figure 4.5 The Use of the COPE Marking Criteria



The statistical analyses carried out to provide evidence for the internal validity of the COPE examination were useful with respect to the information they provided. Regarding the internal validity of the reading paper, statistics pointed to strong and weak areas in the COPE examination. For example, although the reliability of the reading paper is moderate due to the truncated population, it could be further improved. The mean point biserial and facility values revealed that the examination had average difficulty with well-constructed items. In terms of the writing paper, MFR analysis showed that COPE writing criteria were well-constructed, but the markers needed further training.

b) External validation - criterion-related validity of the COPE anchor test

- Teacher Judgments

The first type of criterion-related validity evidence is based on teacher judgments. As a preliminary study, two months before the exam was administered (June 2008), three members of the project group were asked to assess a small sample of students and assign CEFR levels in skills, particularly in reading, listening and writing. This sample was too small to make solid statistical statements about the cut-score but it does allow for moderate approximations. The results, which were based on overall COPE scores, were promising, as shown in Table 4.43. This small-scale study shows a 70.83% agreement (17 students out of 24) between the classifications based on the test performance interpreted using the estimated cut score from the standard setting stage of the project and on the criterion, which is the teacher assessment of student ability.

Table 4.43 COPE CEFR Standard Setting Decision Table June 2008

		COPE (Item Bank)			
		Below B2	B2	Above B2	Total
Criterion (teachers)	Below B2	9	2		11
	B2	5	8		13
	Above B2			0	0
	Total	14	10	0	24

The same study was repeated in January 2009 and June 2009 to further support the validity of the cut-scores set during the standardisation stage of the COPE CEFR linking project. Starting from the beginning of the 2008-2009 academic year, a group of teachers (18 people) were trained in CEFR initially for about three months until the January 2009 COPE and then for another three months until the June 2009 COPE. Before the January 2009 COPE these teachers were asked to assess only 3-4 students in

their classes using the CEFR. The teacher judgments were compared to the COPE results for reading in Table 4.44.

January 2009 results show a 76.19% agreement (32 students out of 42) for reading between the classifications based on the COPE performance and the teacher assessment of student ability.

Table 4.44 COPE Reading CEFR Standard Setting Decision Table January 2009

		COPE Reading (Item Bank)			
		Below B2	B2	Above B2	Total
Criterion (teachers)	Below B2	10	10		20
	B2		22		22
	Above B2			0	0
	Total	10	32	0	42

In January 2010, the sample size in terms of student numbers was even larger with 134 students. In the January 2010 administration of the COPE examination, 10 teachers, who had previously been trained in the CEFR, were asked to make judgments about their students writing ability only. A presented in Table 4.45, the agreement between the classifications based on the test performance interpreted using the estimated cut score (21 is least able B2) from the standard setting stage of the project and on the criterion, which is the teacher assessment of student ability 59.7%. As seen in Table 4.46, the agreement increased considerably in June 2010 with 72.8%.

Table 4.45 COPE Writing CEFR Standard Setting Decision Table January 2010

		Test (Item Bank)			
		Below B2	B2	Above B2	Total
Criterion (teachers)	Below B2	39	19		58
	B2	35	41		76
	Above B2			0	0
	Total	74	60	0	134

Table 4.46 COPE Writing CEFR Standard Setting Decision Table June 2010

		Test (Item Bank)			
		Below B2	B2	Above B2	Total
Criterion (teachers)	Below B2	32	15		47
	B2	13	43		56
	Above B2			0	0
	Total	45	58	0	103

One of the suggestions of the Manual for external validation is the use of teacher judgments to validate the cut scores for the reading and writing papers set during the standardisation stage. As presented above teacher judgments are now embedded in the assessment operations and regularly used to collect criterion-related evidence on the COPE examination. Teacher judgments in this study provided information regarding how accurately the cut scores were set by demonstrating the degree of agreement between teacher's perceptions of learner ability and learners actual test performance. The results showed high agreements between teacher judgments and test scores.

- Correlations (Reading)

The second type of criterion-related validity evidence provided is a result of the correlations between judges' estimates of reading item difficulty established in the

standard setting sessions and the actual difficulty values of those items from the June administration. Table 4.47 indicates that there were acceptable correlations between the judges' estimates and the true difficulty values of the items.

Table 4.47 Correlations between judge estimates and item difficulty values

Correlations	Yes/No Method	Angoff Method
Judge estimates and item difficulty for reading	.51	.49

The Manual suggests the use of correlations between the examination under study and an external criterion measuring the same traits. In this study, correlations were used in a slightly different way. Judges estimates of item difficulty for reading during the standardisation stage was correlated with the actual item difficulty values. Not only did the correlations corroborate the cut scores but they also provided information regarding the performance of the judges. The judges were successful in estimating item difficulty values.

- Comparison with other exams (Reading): FCE and CAE

The third type of evidence comes from comparing the difficulty level of the COPE reading items with FCE and CAE reading items. As the aim was to set COPE at B2 level, two tests were compiled. Test 1 comprised of COPE and FCE (B2 level) reading items. Test 2 consisted of COPE and CAE (C1 level) Reading items. These tests were administered to all students who would sit the June 2009 COPE two weeks later as part of the trialing system. In BUSEL, a trialing system is in place to pilot newly written items so that they can be anchored and stored in the item bank for future use. In order to make the trialing system reliable, the tests are introduced to the students as 'Extra Points Exam'. Throughout their course prior to COPE, students have two achievement tests

and a learning portfolio assessment out of 10. They need to score at least 60% in total to become eligible to sit the COPE. With the help of the Extra Points Exam, students can get an extra 5 points towards eligibility.

Figure 4.6 demonstrates how the COPE reading items are placed on the Rasch logit scale in relation to the FCE items. It can be seen that items on both tests are spread within the ~ +1.00 and ~-1.00 logit scale range. This is also supported in Table 4.48 where the mean facility values of the three examinations are compared. COPE and FCE has very similar facility values with the FCE being slightly easier.

Figure 4.6 COPE and FCE Reading Items

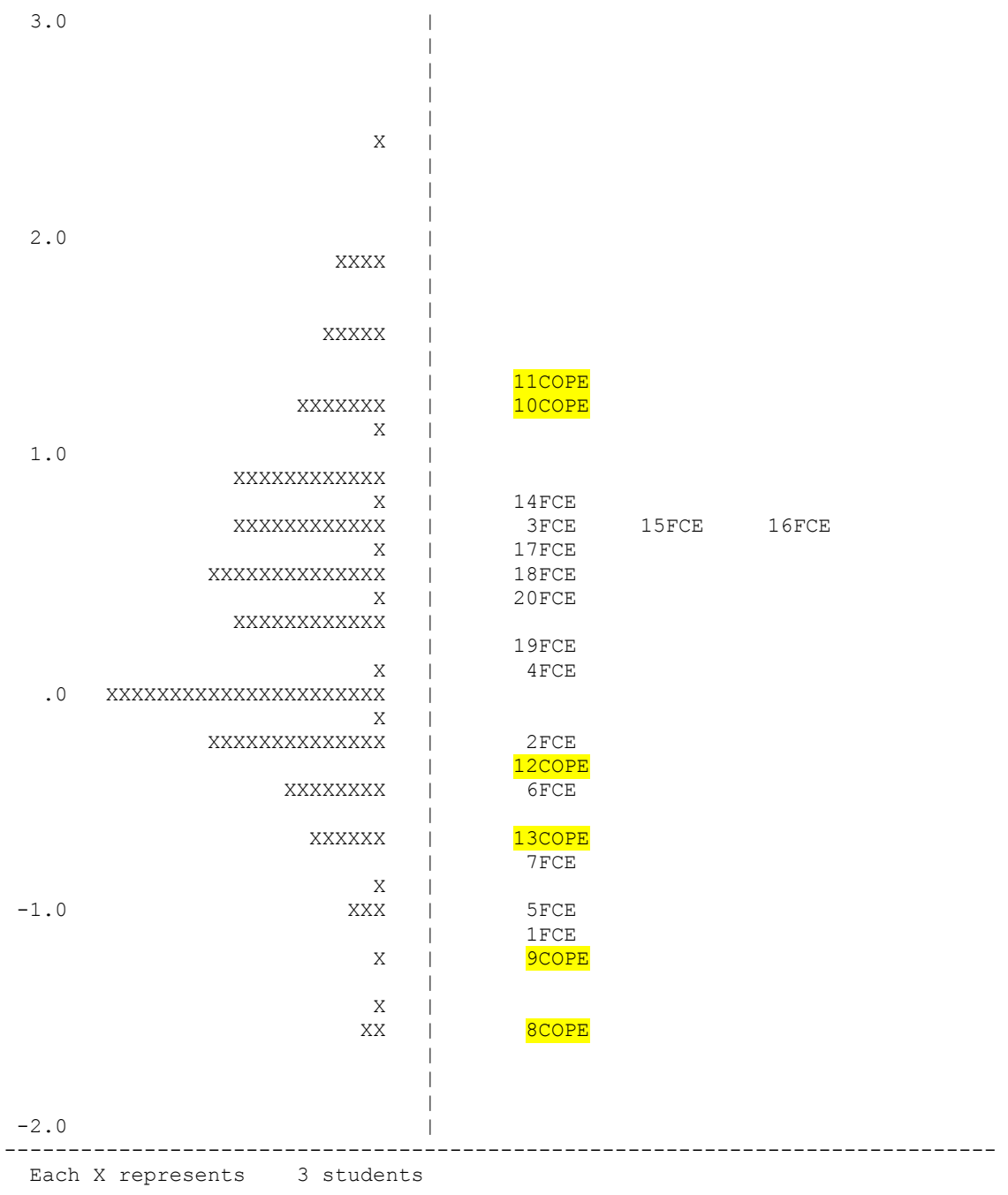
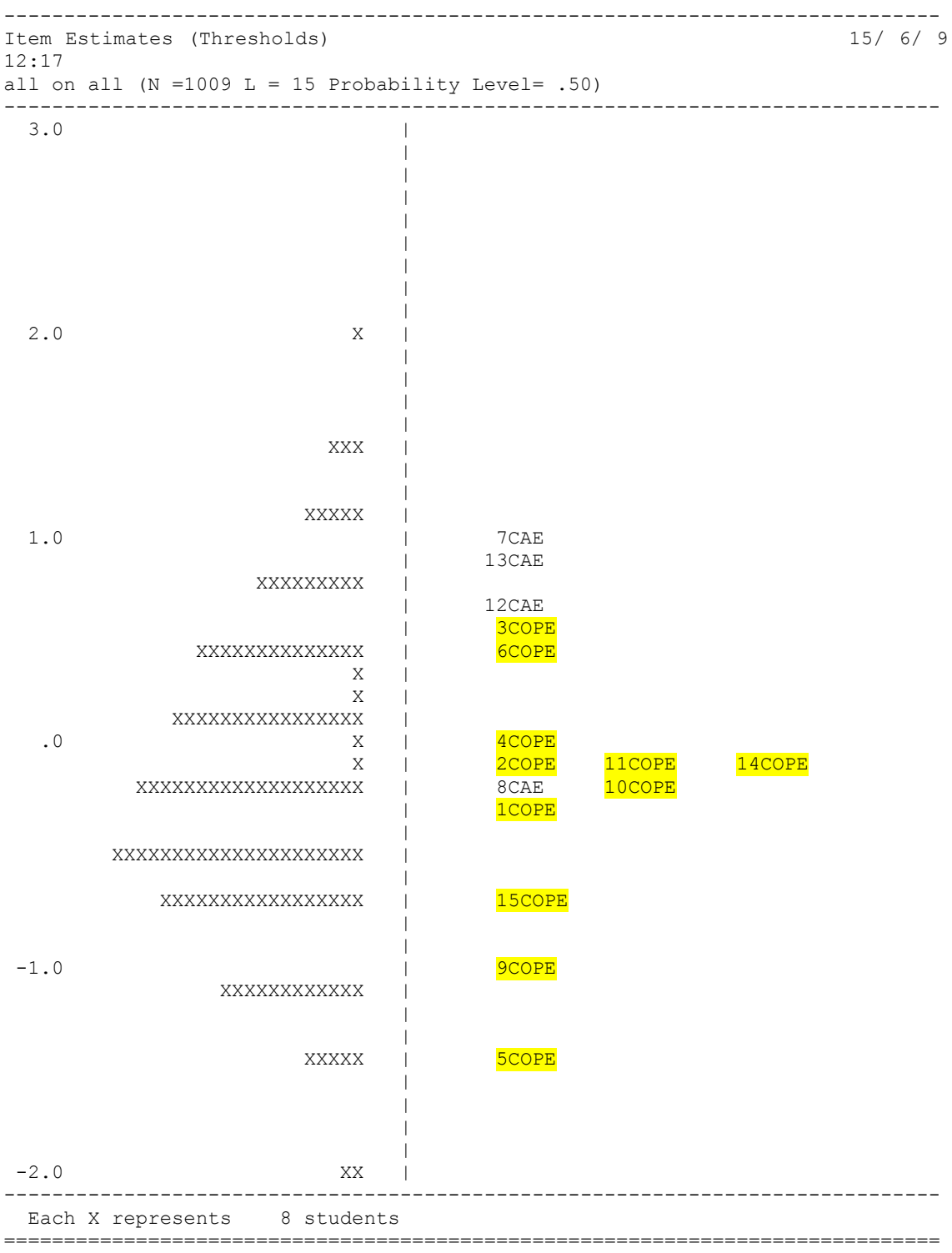


Figure 4.7 in the same way shows the COPE and CAE reading items on a logit scale. With the exception of one item (8CAE), all of the CAE items are placed higher on the continuum than the COPE items. Similarly, in Table 4.49, it can be observed that there

is a clear difference between the mean facility values of the COPE reading items and the CAE items. The CAE items are clearly more difficult than the COPE items.

Figure 4.7 COPE and CAE Reading Items



The mean Facility Values of the three tests show that the COPE items are slightly more difficult than the FCE items whereas the CAE items are considerably more difficult than the COPE items.

Table 4.48 Comparison of Mean Facility Values for COPE, FCE and CAE

	TEST 1		TEST 2	
	COPE	FCE	COPE	CAE
MEAN FACILITY VALUE	59.1	54.1	51.0	31.8

Comparison of COPE, FCE and CAE reading items were also carried out by including items from all three tests in two other 'Extra Points Exam' prior to the January 2010 COPE and June 2010 COPE. This data was analysed using one-way ANOVA, which would reveal whether the differences among the tests were significantly meaningful.

The descriptive statistics presented in Table 4.549 show that there is little meaningful difference between COPE and FCE items whereas the CAE items are more difficult than the others with a difference of approximately 1 logit value. The one-way analysis in Table 4.50, however, suggests that significant difference exists among all three sets of items. This is also supported by the post hoc test in Table 4.51.

Table 4.49 Descriptive statistics for COPE, FCE and CAE reading items-Test 1

Descriptives

SCORE

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
COPE	403	3.5831	1.24185	.06186	3.4615	3.7047	.00	6.00
FCE	403	3.3325	1.65827	.08260	3.1701	3.4949	.00	7.00
CAE	403	2.3474	1.34882	.06719	2.2153	2.4795	.00	6.00
Total	1209	3.0877	1.52265	.04379	3.0018	3.1736	.00	7.00

Table 4.50 One-way ANOVA for COPE, FCE and CAE reading items – Test 1

ANOVA

SCORE

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	343.932	2	171.966	84.416	.000
Within Groups	2456.774	1206	2.037		
Total	2800.706	1208			

Table 4.51 Post hoc tests for COPE, FCE and CAE reading items – Test 1

Multiple Comparisons

SCORE

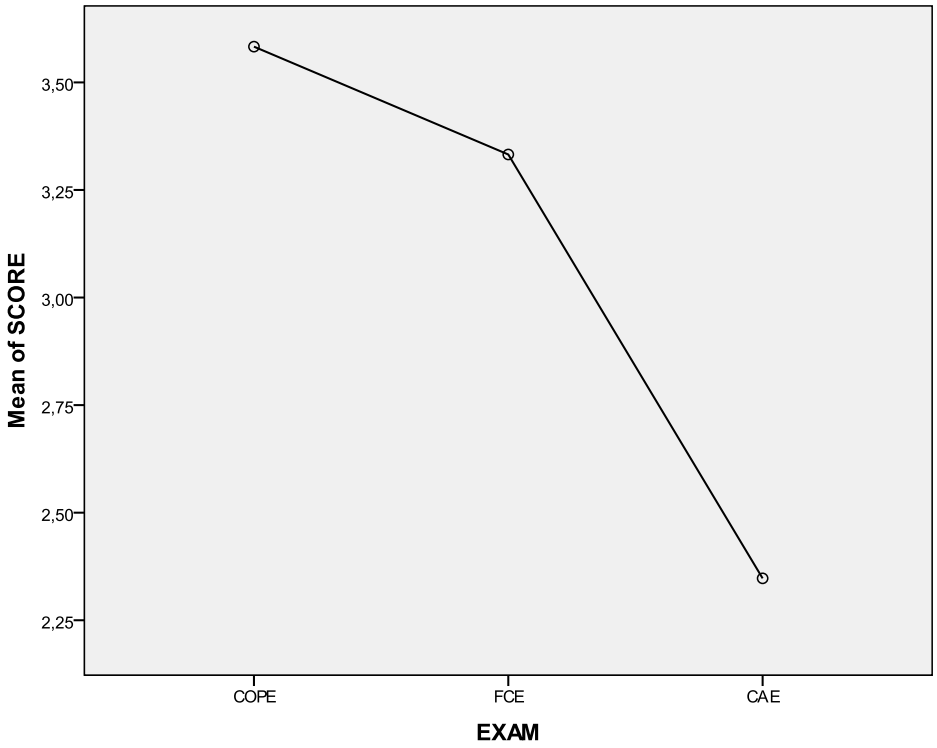
Bonferroni

(I) EXAM	(J) EXAM	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
COPE	FCE	.25062*	.10055	.038	.0096	.4917
	CAE	1.23573*	.10055	.000	.9947	1.4768
FCE	COPE	-.25062*	.10055	.038	-.4917	-.0096
	CAE	.98511*	.10055	.000	.7441	1.2262
CAE	COPE	-1.23573*	.10055	.000	-1.4768	-.9947
	FCE	-.98511*	.10055	.000	-1.2262	-.7441

*. The mean difference is significant at the 0.05 level.

The means plot for the exams, Figure 4.8, also demonstrates the difference between the items. The COPE and FCE items are close to one another on the logit scale whereas this is not the case for CAE items.

Figure 4.8 Means plot for COPE, FCE and CAE reading items – Test 1



The results were slightly different for the second 'Extra Points Exam'. Table 4.52 presenting the descriptive statistics for the three sets of items demonstrate, similar to test 1, little meaningful difference between COPE and FCE.

Table 4.52 Descriptive statistics for COPE, FCE and CAE reading items – Test 2

Descriptives

LOGIT

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
FCE	7	9.4757	.8621	.3259	8.6784	10.2730	8.36	11.01
CAE	4	11.0250	.5720	.2860	10.1148	11.9352	10.24	11.59
COPE	8	9.9463	.7487	.2647	9.3203	10.5722	8.99	11.18
Total	19	10.0000	.9284	.2130	9.5525	10.4475	8.36	11.59

The one-way ANOVA results also show that there is not a significant difference between the sets of items.

Table 4.53 One-Way ANOVA for COPE, FCE and CAE reading items – Test 2

ANOVA

LOGIT

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	6.150	2	3.075	5.253	.018
Within Groups	9.365	16	.585		
Total	15.515	18			

Table 4.54 presents the post hoc test findings, which reveal that the differences between COPE and CAE as well as FCE and CAE are significant whereas there is no significant difference between COPE and FCE items. This can also be clearly seen in Figure 4.9 presenting the means plot for the items from three tests.

Table 4.54 Post hoc tests for COPE, FCE and CAE reading items – Test 2

Multiple Comparisons

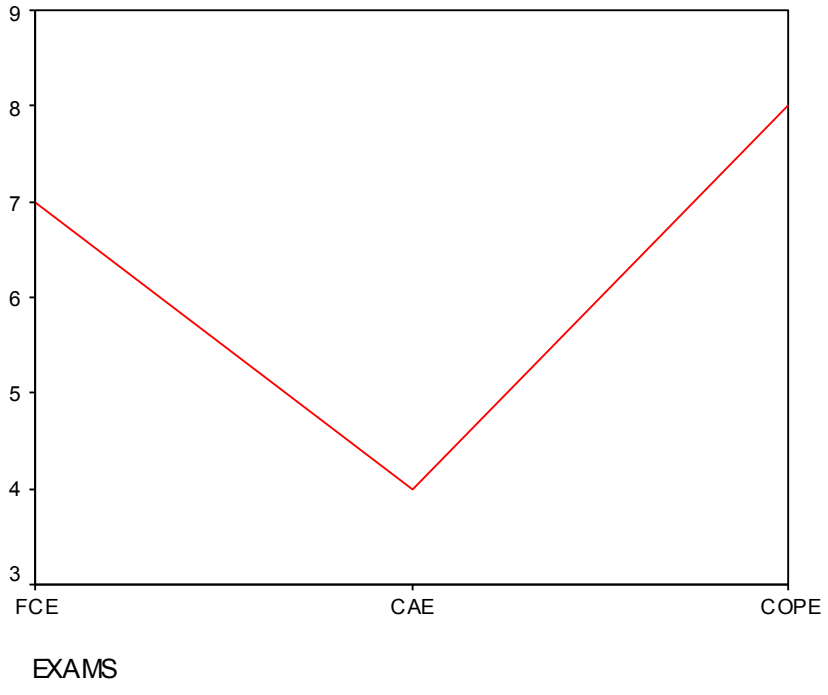
Dependent Variable: LOGIT

Bonferroni

(I) EXAMS	(J) EXAMS	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
FCE	CAE	-1.5493*	.4795	.016	-2.8311	-.2675
	COPE	-.4705	.3960	.756	-1.5289	.5879
CAE	FCE	1.5493*	.4795	.016	.2675	2.8311
	COPE	1.0787	.4685	.105	-.1736	2.3311
COPE	FCE	.4705	.3960	.756	-.5879	1.5289
	CAE	-1.0787	.4685	.105	-2.3311	.1736

*. The mean difference is significant at the .05 level.

Figure 4.9 Means plot for COPE, FCE and CAE reading items – Test 2



As mentioned above, the Manual suggests that the correlation between the examination under study and another test measuring the same abilities can be calculated to

investigate the criterion-related validity of an examination. In this study, rather than using correlation, ANOVA was used because ANOVA allows for the comparison of mean scores or facility values of the two examinations and shows whether the mean differences/similarities are significant or not.

The four sets of data gathered through 'Extra Points Exams' presented here suggest that the results of the calibration attempts were inconclusive. The differences in the results of the one-way ANOVA analysis might be for a number of reasons. One of the reasons could be that the COPE items show variability in terms of difficulty. The same problem may exist with the FCE and CAE items. The second reason might be that the three exams are different in terms of their constructs. For instance, in this study, it was a challenge to find an exam that reflected the construct of the COPE examination. FCE and CAE were used for external validation purposes; however, FCE targets general English and CAE is academic but reflects the C1 level whereas COPE aims to measure academic skills at the B2 level. The results of the comparisons discussed in section 4.5.4.2 revealed that although the COPE reading items were the same in terms of difficulty in the two calibration studies conducted for COPE, FCE and CAE, the results differed greatly. The first one showed that there was no significant difference between the tests, and the second one demonstrated a significant difference between COPE and CAE; and no difference between COPE and FCE. This problem with the differing ANOVA results points to the fact that items reflecting a level can be representative of the top of a level such as the top of the B2 level or the bottom of a particular level. Users of the calibrated items provided by the Council of Europe should be offered more specific information about the items than just the CEFR levels. In other words, it is useful for users to know whether the items provided reflect the top, bottom or the

middle of the level in question so that the users can examine their items and exam more accurately.

4.5.5 Empirical validation stage summary of research findings

The analysis of the interview showed that comparison of the COPE examination with other tests and using teacher judgments to confirm the cut scores established as a result of the standardisation stage did not only help understand the level of the examination but also the level of the students in the Preparatory program, which all contributes to the validity of the examination. In addition, the empirical validation stage gives information as to how the level of the examination can be maintained.

The statistical analyses showed that the procedures followed throughout the empirical validation stage, focusing on scoring and criterion-related validity, provided detailed information about the COPE examination, leading to a stronger validity argument in terms of the validity of the standard setting and the examination in general. However, the calibration of COPE reading items with FCE and CAE highlighted a weakness in the samples provided to accompany the Manual as the calibration was inconclusive, in that it lead to two different conclusions. Precise information regarding the calibrated samples should be provided to users to guide them in their efforts to compare their examinations with external ones in terms of the CEFR.

As discussed earlier in section 4.5.2, the Manual perceives validation as having two aspects; 'internal' and 'external', which is restricted and not in line with the latest developments in validation as indicated by O'Sullivan (2009a, 2009b, 2009c; 2011); and later by O'Sullivan and Weir (2011). Although the Manual suggests that 'internal'

validity is the quality of the test itself and 'external' is the validity of the linkage claim, the analysis tools recommended to achieve these aims form only a limited aspect of validity. 'Internal' validity examines how well the items in a test work and how well the test as a whole function (Classical Item analysis and Item Response theory). It also looks at whether a test measures what it claims to measure (Factor analysis, generalisability). 'External validation', on the other hand, focuses on comparing a test with other tests measuring the same traits, which is only one parameter of criterion-related validity. Therefore, the COPE CEFR project adapted a broader and more recent view of validity by following Weir's validation frameworks that aims to investigate all aspects of a test such as the test takers, cognitive, context, consequential aspects.

The empirical validation stage carried out by the project members was not flawless because, while the external validity (criterion-related validity) evidence accumulated from teacher judgments corroborated the standard setting results, the evidence from the comparison of tests proved to be problematic. Therefore, further research is required in this aspect. Furthermore, In terms of internal validity (scoring validity), the data presented came from one exam and a comparison over different administrations should be made and reported to further support the validity claims. Scoring validity evidence for the examination is reported after every administration and the level as well as the quality of the COPE examination has been sustained over the years because of the item banking system that is in operation.

The results of the empirical validation were promising for a number of reasons. First, the fact that the COPE writing criteria work well, in that the raters are able to clearly distinguish between the bands, as presented in section 4.5.4.2a, was confirmed once

again. Secondly, the reading cut score as a result of the linking project corresponded to the cut score that already existed and that has been implemented over the last eight years, which suggests that the CEFR standard setting will not have a major impact on the pass/fail numbers, as students passing the COPE examination are already at the B2 level, which is deemed acceptable for academic study. This also confirms the intuitively maintained levels initially set for the COPE examination. Thirdly, the teacher judgments, an examinee-centred standard setting method, confirmed the cut scores established at the end of the standard setting.

4.6 Conclusions drawn from Phase 1 of the study

Phase 1 of the research gave initial answers to the research questions. The highlighted boxes in Table 4.55 show the stages of the CEFR linking process and which aspects of validity they contribute to. The ticks (√) indicate the research tools that provided answers to the research questions. Q refers to 'questionnaire', FN stands for 'field notes' and SA is 'statistical analysis'.

Table 4.55 Initial findings resulting from Phase 1

	Familiarisation			Specification	Standardisation			Empirical validation	
	Q	FN	SA	Interview	Q	FN	SA	Interview	SA
RQ1/2a Test taker									
RQ1/2b Context	√	√		√	√	√			
RQ1/2c Cognitive	√	√		√	√	√			
RQ1/2d Scoring		√	√			√	√		√
RQ1/2e Consequential				√					
RQ1/2f Criterion-related								√	√
RQ3a Understanding standards				√	√	√		√	√
RQ3b modifying levels				√		√		√	

Table 4.56 shows that 'test taker' is the only aspect of validity that is not tackled in the CEFR linking process. All other aspects of validity are considered to some degree. Whereas familiarisation and standardisation stages seem to cover cognitive, context and scoring aspects of validity, specification stage only deals with context, cognitive and consequential aspects and empirical validation tackles the scoring and criterion-related aspects of validity. In terms of research question 3, institutional implications, except for familiarisation, all stages of the CEFR linking process contribute to understanding the standards set through the COPE examination and how the level of the examination can be modified to better reflect the intended level, B2 in this case.

The next chapter, Chapter 4, presents Phase 2 of the research that aims to further investigate the research questions and fill in the missing parts of Phase 1.

CHAPTER 5

PHASE 2 – IN-DEPTH ANALYSIS OF THE CEFR LINKING PROCESS

5.1 Introduction

In Phase 1 of the research each stage of the CEFR linking process was explored separately. Phase 2 required participants to look at the experience of CEFR linking as a whole. In this chapter, first, the purpose of Phase 2 is given. Secondly, the data collection procedures are presented. Next comes the data analysis and findings. Finally, conclusions are drawn from this phase.

5.2 Purpose

Data presented in the previous chapter (Chapter 4) showed that the familiarisation stage of the CEFR linking process had a strong focus on scoring validity, in that it required working with a set of criteria, the CEFR scales, and analyzing statistically how the raters used the criteria. Context and cognitive aspects of validity were also in evidence since the participants frequently talked about task demands during familiarisation. Data indicated that the specification stage was restricted to certain aspects of context and cognitive validity as it required an analysis of task and language knowledge parameters as part of the cognitive load as well as gaining a better understanding of the examination. Similar to familiarisation, the standardisation stage was heavily skewed towards scoring validity, as well as context and cognitive validity, as the stage required respondents to apply the standards set through the CEFR to the COPE, requiring an in-depth analysis of the level of the examination. Finally, the empirical validation indicated a significant focus on both scoring validity and criterion-related validity. There are gaps of great importance, which did not emerge from the data in Phase 1. For

example, test taker characteristics and aspects of consequential validity are not considered at any stage of the linking process, and criterion-related validity seems to have come into play only at the empirical validation stage. Therefore, Phase 2 aims to explore these gaps with the help of a questionnaire, focusing on all aspects of validity and addressing all three research questions. The respondents were required to think about the CEFR linking process as a whole and reflect on different aspects of the research questions.

5.3 Data collection

The questionnaire (Appendix 3L) was designed based on Weir's validation framework where different aspects of validity such as context validity or scoring validity, together with their parameters were included. For instance, the parameters of scoring validity such as item analysis, reliability, error of measurement, were listed. Each parameter formed a question in the questionnaire and was explored in relation to each stage of the CEFR linking process for writing and reading separately, where parameters were not common. In addition to questions related specifically to a validation model, Research Questions 3a and 3b, which focused on issues regarding the level of the examination, were also explored through the questionnaire for reading and writing separately. A more complete overview of the questionnaire is available in section 3.7.2.

The questionnaire was administered to ten respondents after most of them had experienced each stage of the linking process, viz. familiarisation, specification, standardisation and empirical validation, though at a time when the project was still under way. The standardisation stage was repeated for reasons explored in Chapter 4 and collection of further data for empirical validation was still continuing. Only ten

project members from the original fourteen were available to answer the questionnaire because by that time two project members had stopped working in BUSEL, one had left the project and another one was on leave of absence. Of these ten, one did not taken part in the specifications stage and another did not participate in the writing standardisation. Therefore, for these two stages, there were only nine responses to the questionnaire. The questionnaire was conducted towards the end of the project as it was crucial for them to have seen the whole process so as to reflect on it with a better understanding of what CEFR linking entailed. The participants were asked to look back over a period of almost two years while answering the questionnaire. Therefore, in cases where the participants could not remember the sessions, support was available through providing the session overviews and a copy of the tasks carried out in those sessions, where required.

5.4 Data analysis and findings

As Robson (1993) suggests, a simple means of exploring many quantitative data sets is frequency distributions. Therefore, bar charts are used to display data in this chapter and the visual nature of such histograms eases interpretation.

For each aspect of validity viz. test taker, context, cognitive, scoring, criterion-related and consequential validity as well as Research Question 3, which looks at level issues, there is a bar chart that demonstrates the extent to which aspects of validity or issues regarding level are dealt with at each stage of the CEFR linking process. The numbers 0 to 10 on the vertical axis represent the people who answered the questionnaire. The phrases on the horizontal axis are the parameters of a certain validity aspect. The color coding as explained on the right hand side of the bar charts are the stages of the CEFR linking process.

5.4.1 Test taker characteristics

Test taker characteristics are dealt with under three categories; physical/physiological needs, psychological characteristics, and experiential characteristics. Only two out of ten participants thought psychological and experiential characteristics of test takers were taken into consideration for reading and writing at specification and standardisation stages (See Appendix 5A for the results of the questionnaire). The number of responses seems to be too low to lead to any concrete conclusions, but it might suggest that test taker characteristics seem to have almost no place in the CEFR linking process.

5.4.2 Context validity

Weir (1993) sees context as having a prominent role in determining language ability and suggests that if an examination reflects real-life tasks that are based on contextually appropriate conditions and operations, then it would be easier to state what a learner can do. The CEFR also has an action-oriented approach where learners of a language are expected to carry out tasks in real-life situations (Council of Europe, 2001) and these are reflected in the CEFR levels and scales. In part two of the questionnaire, the respondents were asked to indicate which of the areas under context validity they took into consideration in relation to their appropriacy to the target context and at what stage of the linking process these were considered. These areas, explored through 19 questions in the questionnaire, involved task setting parameters (purpose, format, criteria, etc.); task demands (discourse mode, text length, content knowledge, etc.); and, administration setting (security, uniformity and conditions). In the sub-sections below, the extent to which context validity parameters, i.e. setting, task demands and administration, are emphasized throughout the CEFR linking process according to the

respondents who took part in such a study in the BUSEL context is presented for reading and writing separately.

5.4.2.1 Context validity issues for reading

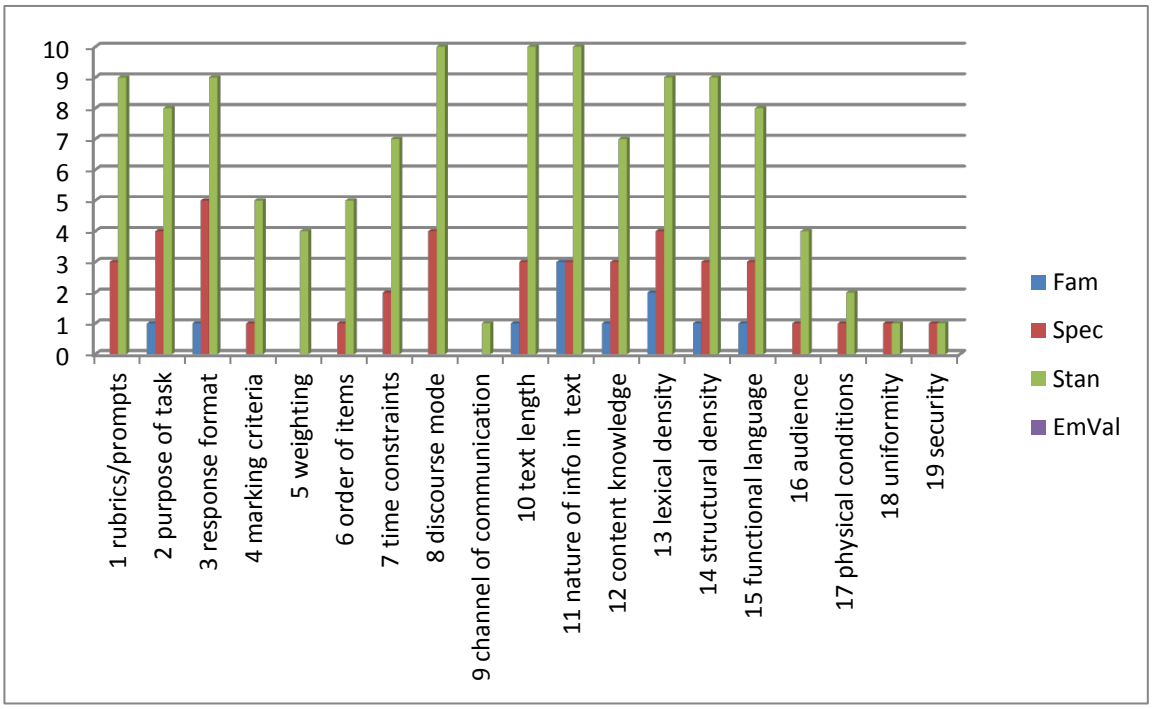
In Figure 5.1, the blue bars indicate the aspects of validity shown by respondents as being considered during the familiarisation stage of the CEFR linking process. The blue bars suggest that issues regarding context validity do not seem to have been taken into consideration by the participants at the familiarisation stage. The nature of information in reading passages (question 11 in the graph), mentioned by 3 out of 10 respondents, seems to have been focused on the most in terms of reading, as text features such as abstractness or concreteness of input and lexical/grammatical density contribute to determining at what CEFR level a person can cope with a certain text.

The results also suggest that at the specification stage, parameters of context validity, such as purpose of task indicated by 4 respondents, response format by 5, discourse mode by 4, or lexical density by 4, received more attention than they did in the other stages of the linking process.

Aspects of context validity were mostly tackled at the standardisation stage for reading. All the respondents drew attention to discourse mode (the genre or the text type), text length and nature of information in the input material. All these aspects are related to task demands, which lies at the core of standard setting, in that, by looking at the task demands participants try to assign levels to reading items.

Context validity was not tackled at all at the empirical validation stage as no purple bars representing empirical validation can be seen in Figure 5.1. However, valid conclusions cannot be drawn regarding the empirical validation stage as with the exception of two, project members were not involved in this stage. Comments regarding this are presented in the implications and conclusions section of this chapter (Section 5.5).

Figure 5.1 Context validity for reading



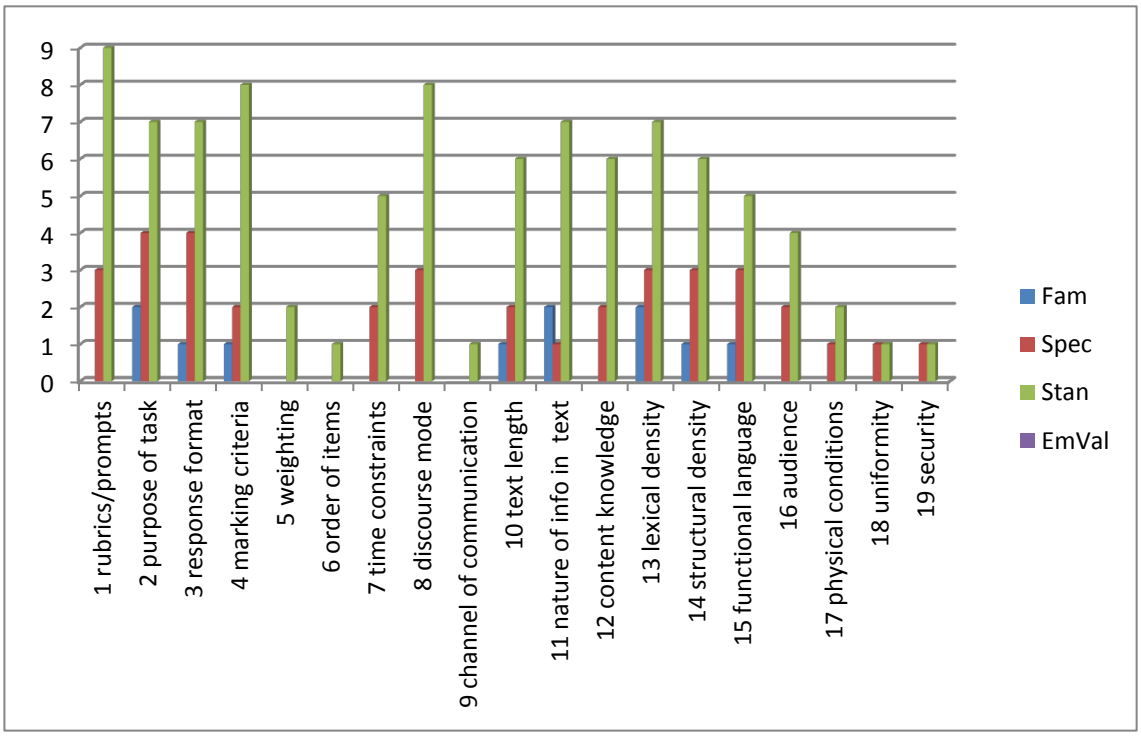
The data shows that certain aspects of context validity appear to be rarely tackled in the CEFR linking process for reading, viz. physical conditions, uniformity of administration and security, which are related to the administration of the examination (17, 18 &19 in the graph). In this respect, they may not play a role in determining the level of a test. However, one of the aims of the specification chapter (Chapter 4) in the Manual is "to contribute to increasing awareness among developers of quality language examinations" (Council of Europe, 2003: 29) and quality also encompasses the quality of

administration of a test, as I2 at the specification stage interview pointed to this weakness of the Manual (Section 4.3.5).

5.4.2.2 Context validity issues for writing

Figure 5.2 suggests similar findings for writing to those of reading in terms of familiarisation, specification and empirical validation stages. In addition, issues regarding test administration were also relevant for writing. Figure 5.2 attests that aspects of context validity for writing received more and more emphasis as the project moved to the standardisation stage, but had no importance at all at the empirical validation stage. Aspects such as prompts (9 out of 9), marking criteria (8 out of 9) and discourse mode (8 out of 9) were particularly important at the standardisation stage. It seems that in determining the level of a written performance, the task, that is the prompt, the marking criteria – how the samples are marked (the CEFR scales in this case), and the discourse mode (the genre) can be expected to play a significant role. In addition, parameters of context validity such as text length, nature of information in the output text, content knowledge, lexical and structural density also had an important place in the standardisation stage. These parameters are also specified in the task difficulty criteria proposed by O'Sullivan and Weir (2011) in assessing the level of a written task, which not only demonstrates why these parameters in particular have a place in standard setting but also supports the idea behind the proposed criteria.

Figure 5.2 Context validity for writing



5.4.3 Cognitive validity

Test designers are ideally required to adapt a language theory such as Bachman and Palmer's model of language ability (1996) according to their needs, and design an examination based on this model although many cannot due to the shortcomings of language theories regarding skills in particular as discussed in Chapter 2 of this research. By doing so, designers make claims regarding what they test through that examination. For instance, is the test a test of careful reading or are sub-skills and strategies of reading tested separately? Cognitive validity involves ensuring that an examination actually measures what it claims to measure in terms of the cognitive processing required to respond to test tasks. In his validation framework, Weir (2005a) included different models; mainly Urquhart and Weir (1998) for reading and Grabe and Kaplan (1996) for writing. In the questionnaire, the respondents were provided with the components of these models such as word recognition, type of reading/writing,

strategies, textual knowledge or sociolinguistic knowledge in 14 questions for reading and 15 for writing and were asked to indicate at which stage of the CEFR linking process they considered these components for reading and writing.

5.4.3.1 Cognitive validity issues for reading

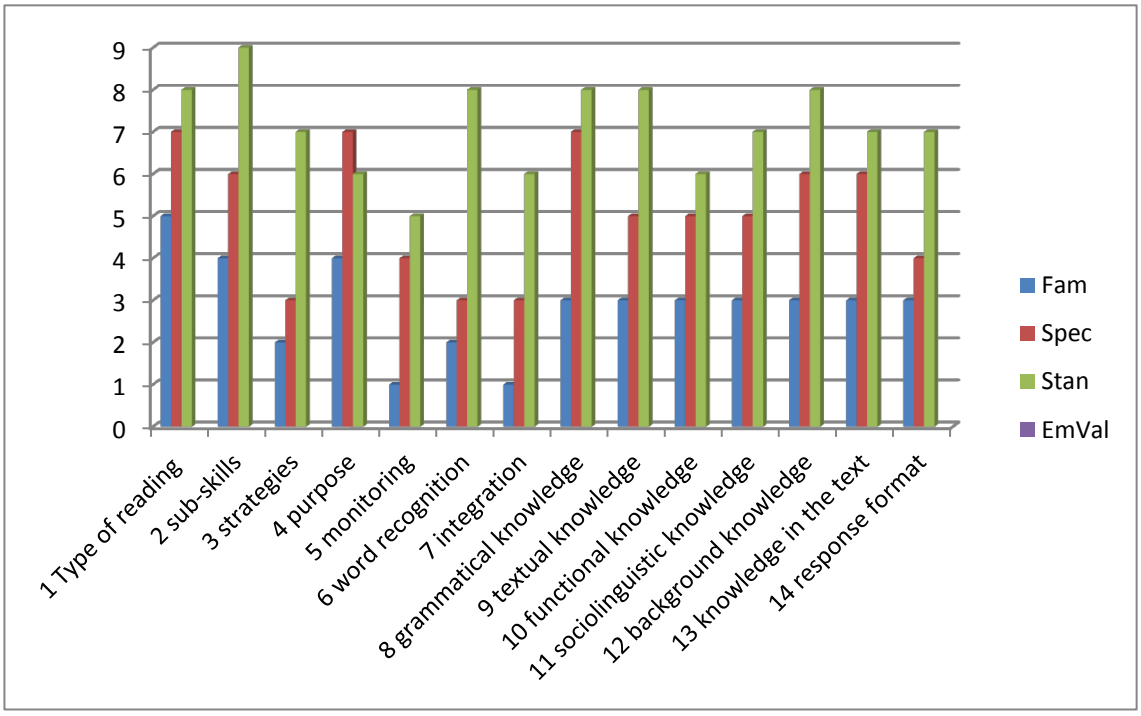
In terms of cognitive validity, a different profile from that of context validity emerges for reading at least at familiarisation and specification stages. At the familiarisation stage, while working on the CEFR descriptors and scales, the respondents mostly paid attention to the type of reading (5 out of 10), sub-skills (4 out of 10), and purpose of a reading task (4 out of 10). This seems like an expected result as the CEFR descriptors focus on these elements. For instance, the following B2 level descriptor taken from the Reading for Orientation scale (Council of Europe, 2001: 70) specifies the type of reading (scanning) and the purpose of task (locate relevant information or identify the content).

> *Can scan quickly through long and complex texts, locating relevant details.*
> *Can quickly identify the content and relevance of news items, articles and reports on a wide range of professional topics, deciding whether closer study is worthwhile.*

At the specification stage represented by the red bars in Figure 5.3, in addition to the elements highlighted for the familiarisation stage, grammatical density in the text was also seen to be important for respondents as it was indicated by 7 out of 9 respondents. This might be due to the need to take into account the linguistic difficulty of a text in order to assign a level to it.

Standardisation required the respondents to actually answer the reading items and consider how they answered them. This, as demonstrated in Figure 5.3 through the frequency of the green bars, called for consideration of all parameters of cognitive validity though to relatively different degrees. This is typified by the fact that sub-skills were indicated by all respondents as opposed to monitoring one's own reading, which was indicated by 5 out of 10 respondents. This is an inevitable result at this stage because, as explained earlier on, cognitive validity is all about the theory or model of language ability, particularly reading in this case. The model chosen must be capable of describing the cognitive processes one goes through while answering the reading items of a given test. Because the participants were actually doing the test and trying to verbalize or at least describe in their minds what processes they used, all aspects of cognitive validity were in action at this stage.

Figure 5.3 Cognitive validity for reading

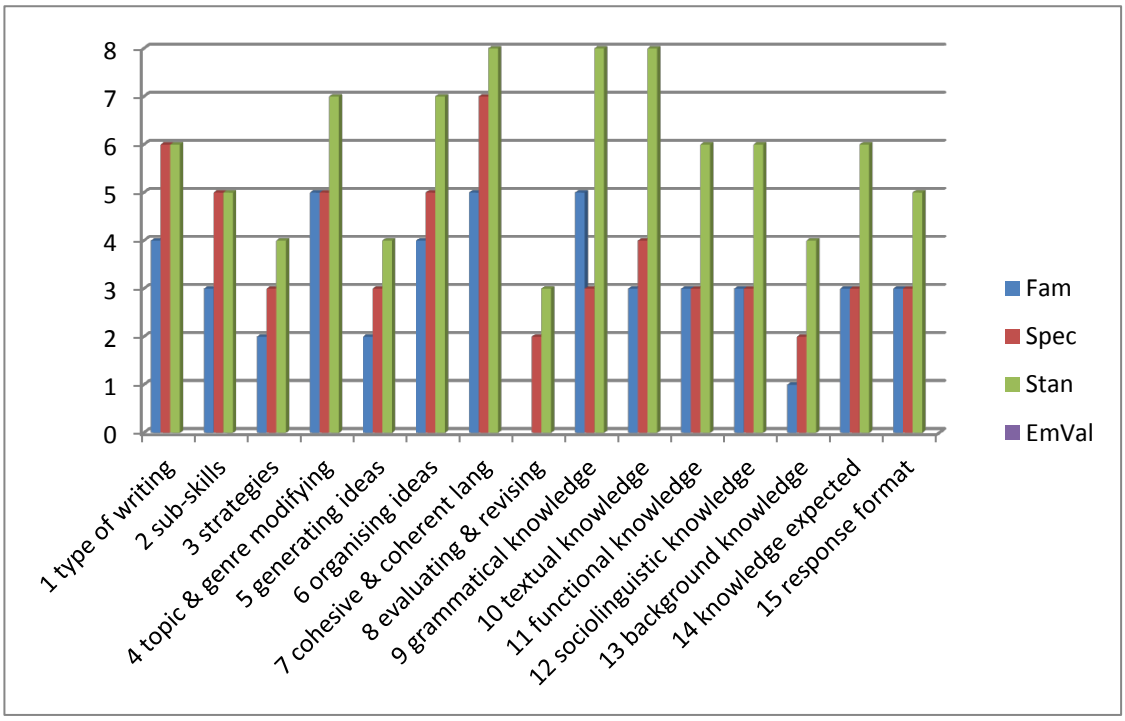


Elements of cognitive validity, on the other hand, were not addressed at the empirical validation stage as none of the parameters of cognitive validity were specified by the respondents. This might be for two reasons: the empirical validation stage proposed by the Manual is too limited in the sense that the Manual's approach to validation is not as broad a view of validation as the Weir model suggests, thus not taking into consideration all aspects of validity; or this aspect of validity was not investigated by the members of the project within the time period of the study.

5.4.3.2 Cognitive validity issues for writing

For the skill of writing, elements of cognitive validity again play a significant role in understanding the requirements of a certain CEFR level at the familiarisation stage. What is not taken into consideration at this stage is 'evaluating and revising' labeled as number 8 in Figure 5.4. An explanation for this could be that 'evaluating and revising' are sub-skills that cannot be observed in a piece of written performance in the context of CEFR linking studies where people work with end-products. Even if a student had undertaken evaluation and revision while producing a piece of writing as part of a test, this cannot be observed unless there are physical indications such as crossing out chunks, sentences or sections.

Figure 5.4 Cognitive validity for writing



At the specification stage, the type of text (genre) (number 1 in the graph) and particularly textual features such as putting ideas into appropriate, cohesive and coherent language (number 7), seem to have been important as indicated by 6 to 7 respondents out of 9. Standardisation represented in green in the graph is again the stage where all parameters of cognitive validity were considered at different levels while assigning levels to sample written performances with aspects such as *putting ideas into appropriate language* (number 7), *grammatical knowledge* (number 9) and *textual knowledge* (number 10), as part of the cognitive load imposed by a task, used by all the respondents.

5.4.4 Scoring validity

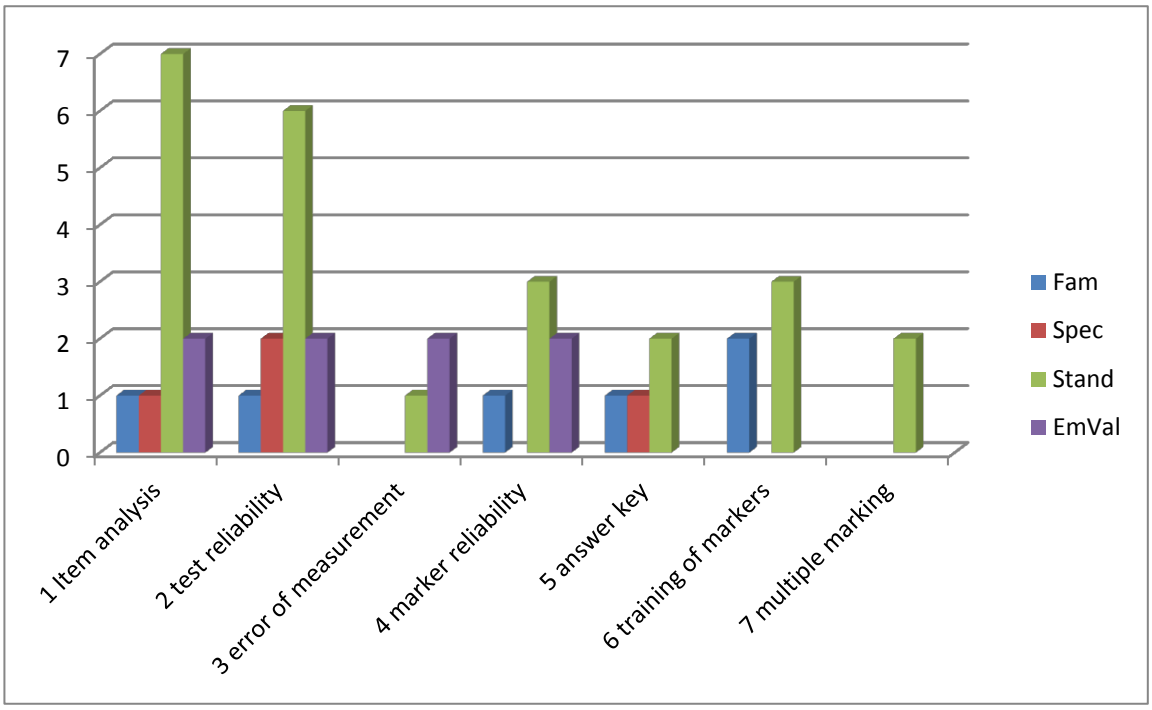
Parameters of scoring validity proposed by Weir (2005a) such as marker reliability, standardisation, item analysis or reliability have the potential to have an impact on the overall reliability of an examination. Scoring validity entails providing evidence as to

the quality of the marking carried out and the test itself or how well the markers rated the writing papers. The participants who answered the questionnaire were asked to indicate whether they took these elements into consideration at any stage of the CEFR linking process for reading and writing. The questionnaire had 7 questions for reading and 10 for writing.

5.4.4.1 Scoring validity issues for reading

A similar pattern emerges to those of the validity types tackled in the preceding sections. Elements of scoring validity become important at the standardisation stage. Particularly item analysis and test reliability become prominent at this stage since the participants were presented with item analysis results during standard setting, which might affect their judgments. Providing such data to participants especially after round 1 of their judgments is common practice in a standard setting event. The aim here is to help them evaluate their judgments regarding the difficulty of reading items.

Figure 5.5 Scoring validity for reading



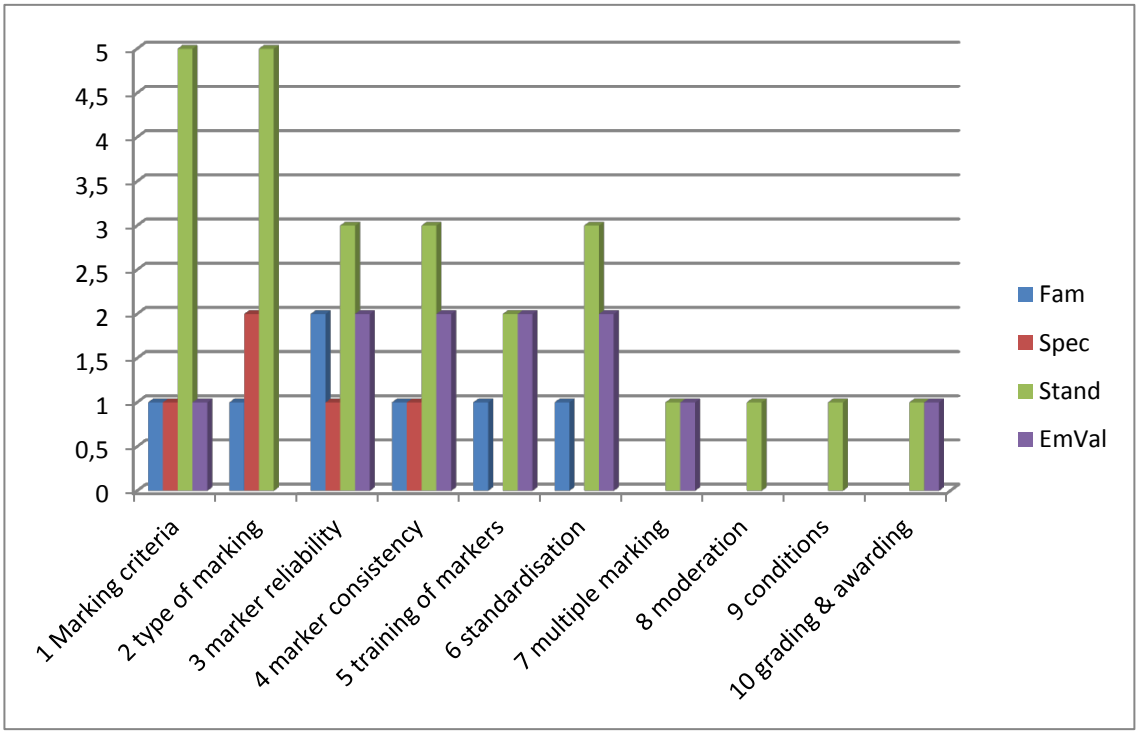
It is worth pointing out once again that only two people filled in the section about the empirical validation stage. The others indicated that they had not taken part in the empirical validation stage and, therefore, did not know what it involved. Both respondents who filled in the empirical validation stage felt that at this stage item analysis, test reliability, error of measurement and marker reliability have a significant place. However, these cannot be generalized as the number of respondents involved in this part of the process is too low.

5.4.4.2 Scoring validity issues for writing

As was the case with reading, for writing, the standardisation stage appears to be significant in terms of scoring validity. Marking criteria and type of marking (holistic vs. analytical) (5 out of 9 people) especially have a place in standardisation as participants try to make connections with the CEFR levels assigned to written scripts and the BUSEL criterion, which is holistic, as opposed to the Written assessment grid (a

grid that is provided in the Manual to be used as the criteria while making judgments about written performances) which encompasses holistic and analytical criteria for marking.

Figure 5.6 Scoring validity for writing



5.4.5 Consequential validity

With respect to consequential validity, investigated through three questions on differential validity, washback and effects on society, only two respondents made comments about the CEFR linking stages for the reasons expressed in Section 5.4.4. This may also be because of the fact that the respondents had a more direct interaction with the other areas explored in the questionnaire, in that, they were personally involved in the familiarisation, specification and standardisation stages of the project. The results for consequential validity appear to be insignificant; however, they both indicated that issues regarding differential validity and washback in classroom or workplace were

considered for reading and writing only at the standardisation stage. The results were restricted to the discussions on these issues without any actual bias analysis or washback study findings.

5.4.6 Criterion-related validity

In terms of criterion-related validity, with 4 questions in the questionnaire, what stands out is that four out of ten people believed that comparison with different versions of the same test is an area considered for the reading paper and two out of 10 felt the same way with regard to the writing paper at the standardisation stage. However, this may not be meaningful due to the low number of respondents. In addition, again at the standardisation stage, three out of ten people stated that comparison with other tests was taken into account for reading and writing papers. Like consequential validity, the respondents did not have direct involvement in the empirical stage and were thus not in a position to make comments about this stage.

5.4.7 Implications of the linking process

Questions regarding implications of the CEFR linking process were prepared independently of Weir's validation framework. These five questions explore institutional implications resulting from such a process.

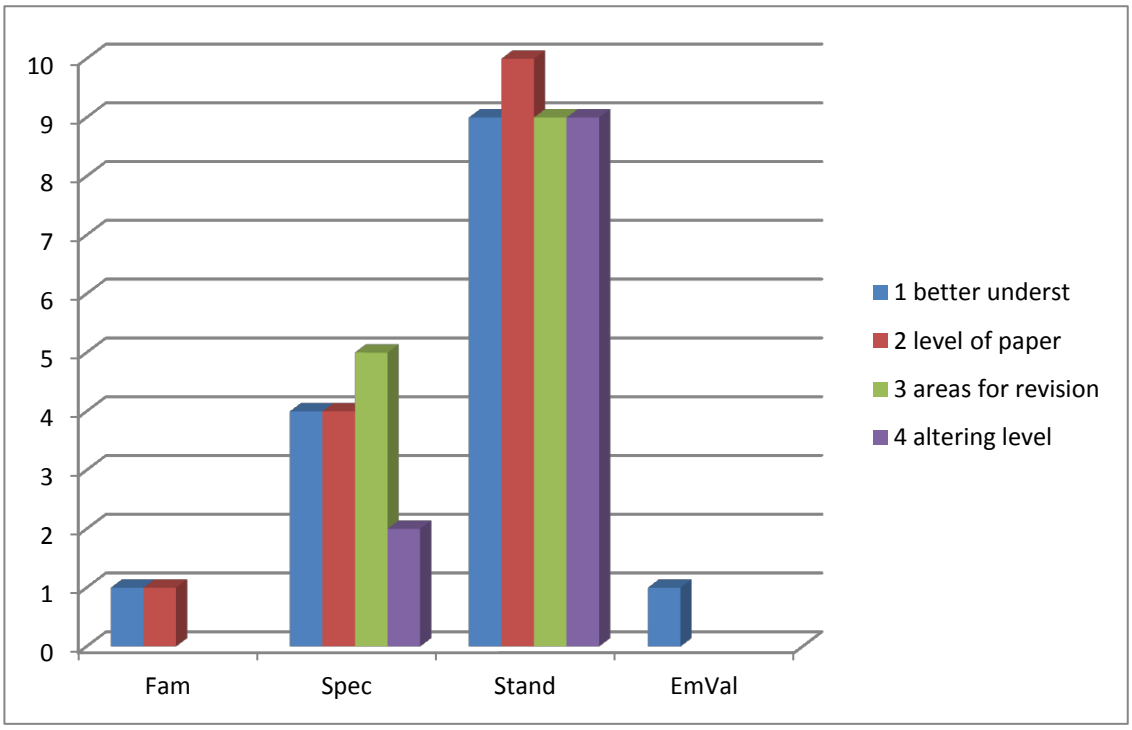
5.4.7.1 Implications for reading

Regarding the implications of the CEFR linking process on the COPE reading paper, the specification stage contributed, indicated by 5 out of 10 respondents, to a better understanding of what it measures, represented in blue and its level in red in Figure 5.7. This stage also pointed to areas for revision in the reading paper (green) indicated by 6

out of 10 respondents. This issue was in fact spotted by O'Sullivan (2009a, 2009b, 2009c), who incorporated a 'critical review' phase to the CEFR linking process in the City and Guilds project. The critical review phase took place before the standardisation stage and required a group of experts in the CEFR to critically analyse the exam to be linked. The aim of the analysis was to ensure that the exam was at the level claimed in terms of tasks and texts. If not, revisions were made to the exam before moving on to the standardisation stage.

It was mainly the standardisation stage, as indicated by 9 to 10 respondents, that had the greatest contribution with respect to the four areas viz. better understanding of what the exam measures, its level, areas for revision and areas to alter its level, whereas the familiarisation and empirical validation stages were perceived to have had almost no value. The familiarisation stage might have been too abstract for the respondents at the time and the empirical validation did not require their direct involvement. These might be the reasons why they thought familiarisation and empirical validation did not have a role in understanding the COPE examination and its level. This issue is further discussed in section 5.5 of this chapter.

Figure 5.7 Implications for Reading



Additionally, the last section of the questionnaire also asked for the participants' opinions of the suitability of the level of the reading paper for its purpose, which is academic study. The standardisation stage set a cut score of 21 out of 35 for a least able B2 candidate and 9 out of 10 respondents were satisfied with this level for academic study considering the competence level of a B2 student. However, the best way to investigate this would be to conduct further research whereby the language performance of students who pass with 21 in their academic studies is evaluated.

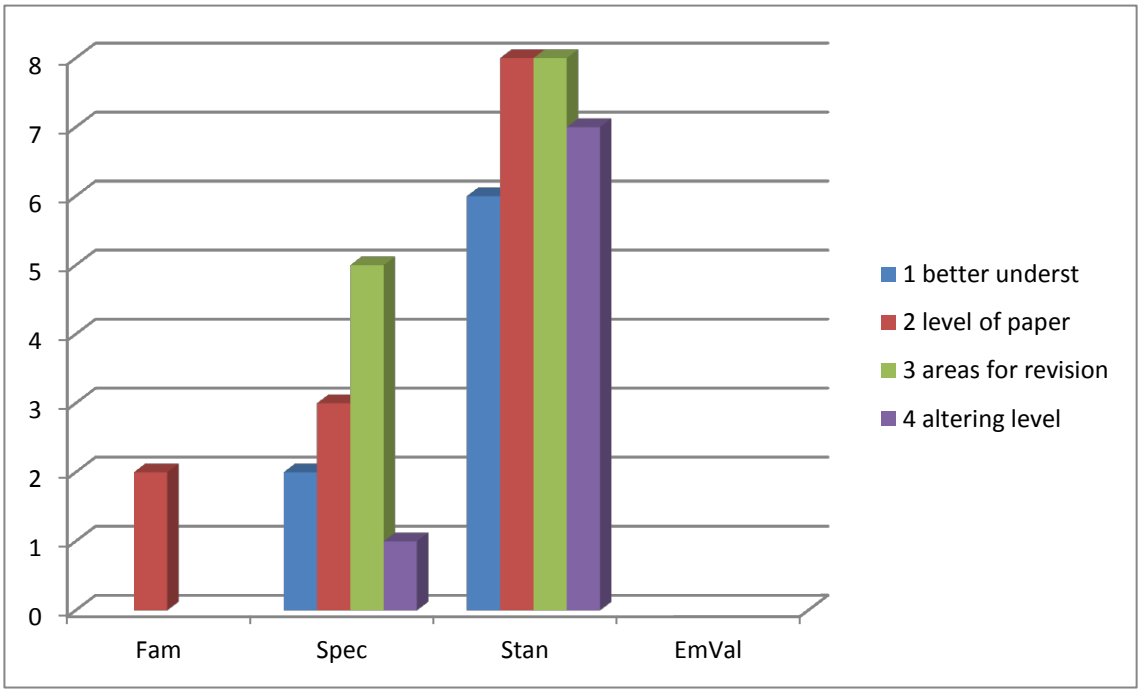
5.4.7.2 Implications for writing

In terms of the writing paper, the contribution of the specification stage was limited to identifying areas for revision. The standardisation stage on the other hand contributed the most to the writing paper. It helped better understand what the writing paper measured (as suggested by 6 out of 10), its level (8 out of 10), areas for revision (8 out

of 10) and highlighted possible areas that could be focused on to alter the level of the paper (7 out of 10).

As for the suitability of the level of the writing paper, 7 out of 10 respondents considered a score of 18 out of 30 that approximately corresponds to a least able B2 performance adequate for academic study.

Figure 5.8 Implications for Writing



5.5 Conclusions drawn from Phase 2 of the study

The challenge of Phase 2 of the research was to measure the performance of the CEFR linking process relative to a theory of validation because the Manual does not have a validation model to guide users, even in its final version (Council of Europe, 2009). One might argue that it is not within the domain or aims of linking to the CEFR or any external benchmark to validate the examination in question as the Manual suggests that the exam to be linked has to be valid as a prerequisite. However, validating the link or

the linking process would seem essential to be able to make meaningful claims about such a linkage simply because linking or setting standards is an important and integral part of any validation procedure. The Manual does suggest that evidence regarding the validity of the linking process, procedural validity, should be collected throughout the process. However, collecting evidence needs to be principled; therefore, it definitely requires a validation framework, ideally based on a validation model. Moreover, if the users of the Manual are asked to provide evidence regarding the internal and external validity of the exam analysed, then besides providing the tools to do so, a validation theory could have also been recommended or at least the suggestion to use one could have been made.

In the BUSEL context, adopting Weir's validation framework to the CEFR linking process led to thought-provoking results for reading and writing. The discussion of the results from this point on will be given in the order in which the aspects of validity are presented in Weir's framework.

It appears that test taker characteristics are not highlighted in the CEFR linking process, as indicated in 5.4.1, perhaps because of the assumption that any organization should take these into consideration and this is at least claimed to be common practice. According to two respondents, test taker characteristics (physical/physiological, psychological and experiential) are tackled at the specification and standardisation stages. However, more emphasis could actually be given to these issues because they cannot easily be controlled but have a direct impact on student performance. For example, knowing the candidature well allows test writers to choose topics that would appeal to the test taker profile and prevent bias towards a certain group of learners. At

least through the specification stage, attention can be drawn to test taker characteristics by including a detailed section on the test taker.

Test taker characteristics are crucial in linking studies for three reasons. Firstly, if the participants of a CEFR linking study do not have a clear picture of test taker characteristics, they may not be able to decide whether an item or a task is appropriate for the level or the target situation. Secondly, most standard setting methods such as the Angoff and Yes/No methods require a good conceptualization of the target test taker. Without knowing the test taker characteristics, it would be, for instance, impossible to define 'the least competent' learner. Finally, in line with the second argument regarding the necessity of taking test taker characteristics into consideration in linking, particularly at the standardisation stage, a meaningful cut score will not be properly established for the typical candidature of an examination.

Even though context validity is meant to highlight the rationale behind the design of a test and its administration procedures, the findings of the questionnaire indicated that it is mostly likely to be considered at the standardisation stage both for reading and writing. It was not prevalent in the other stages, which appears to be a lack. Context validity may in fact come in to play at all stages. For instance, at the familiarisation stage those undertaking a linking study could, first of all, question the appropriateness of the CEFR to their context, that is, how they aim to measure what they want to measure. For instance, CEFR has an action-oriented approach to language learning, which may not be suitable for all tests. The participants of a linking study should demonstrate a full knowledge of the test as well as the CEFR and this knowledge should be ensured through a variety of activities, such as the quizzes used in this study, before

proceeding onto the other stages. At the specification stage, the design of the examination could be analysed in some detail. For instance, the specification forms could have a section where the users are required to rationalize their choice of tasks and the suitability of these tasks for the test takers and the intended attainments to be measured. Finally, at the empirical validation stage, the users of the Manual could be encouraged to gather evidence regarding the context validity of the examination they work on through, for instance, the use of a questionnaire given to test takers or the judges taking part in the study.

Cognitive validity and context validity have several parameters in common. The preceding paragraph indicates that context validity focuses on the design of a test. Part of the design requires test writers to specify the task demands, such as the nature of information or lexical, structural and functional knowledge required to cope with a given task. Similarly, cognitive validity highlights the internal processes taking place in test takers' minds while doing that task. This also includes cognitive load, i.e., language or content knowledge. Evidence on cognitive validity, in fact, tries to describe what actually happens in reality as opposed to what is aimed at, at least in part, through context validity. Therefore, in an exercise such as this, where we explore the extent to which people take into account these aspects of a test when making linking-related decisions, one would expect similar results for cognitive validity to those of context validity as linking-related decisions involve discussions surrounding task demands or cognitive load, which is a parameter common to both context and cognitive validity. However, it is observed that parameters of cognitive validity are considered more throughout the CEFR linking process, not surprisingly at the standardisation stage in particular as presented in section 5.4.3. Throughout the linking process, the participants

of the project tried to verbalize or at least understand what could be going on in test takers' minds working on the reading or writing tasks. Similar to context validity, more effort could be spent to encourage users of the Manual to think about the cognitive validity and accumulate evidence that there is a match between the cognitive and context sides. For example, test specifications outlining the parameters of context validity for a given test could be analysed to see how well the items reflect the parameters during standardisation. This issue is further discussed in the concluding chapter (Chapter 7). However, the reason why the linking process does not force users of the Manual to focus more on context and particularly cognitive requirements of examination tasks might result from the fact that CEFR ability levels are not sufficiently defined.

The results regarding aspects of cognitive validity also suggested that the CEFR had some theoretical underpinnings. Even though CEFR is criticized for lacking theory (Alderson, 2007: Fulcher, 2004a; 2004b, Huhta, et al, 2002; Little, 2007; Weir, 2005b), in defining levels of language proficiency, it inevitably made use of components of language ability as defined by different models. It seems that the results displayed in Figure 5.3 are evidence of this. In other words, it was indicated above that Weir (2005a) designed his validation framework for reading based on Urquhart and Weir's (1998) model for reading ability. Components of this model in the reading validation framework were included in the questionnaire and the respondents ticked several of these as being considered for the linking of the COPE reading paper. This might mean that one or more theories of reading form the basis of the CEFR and thus were considered in the linking process. This, in fact, seems to contradict the criticisms made for the CEFR that it lacks theory (Alderson et al., 2004; Weir, 2005b, Fulcher, 2004a;

2004b; Hulstijin, 2007). However, this finding does not demonstrate which theories the CEFR reflects. The questionnaire only investigated whether certain common aspects of reading models were discussed in the linking process. It did not require respondents to reflect on a given reading model or whether the aspects presented through the CEFR represented good reading.

Results showed that only certain aspects of scoring validity such as item analysis and reliability for reading and marking criteria for writing, were prominent at the standardisation stage. Participants look at item analysis results for instance while making judgments about reading items in standard setting. As for writing, they consider the criteria used for marking, be it the actual criteria of the test or the CEFR scales during standard setting. However, scoring validity should have a significant place in a linking process, notably in two ways. One of these is the scoring validity of the examination to be linked. At the specification stage, aspects of scoring validity could be questioned in some detail so that a message as to the significance of scoring can be sent to the Manual users. This could have an awareness-raising purpose. Secondly, the scoring validity of the linking process itself is also of utmost importance as it provides evidence of the quality of the process carried out and helps validate it. Therefore, unlike the case of the Manual and its final version where most of scoring validation is left to the empirical validation stage, scoring validity could come in at all stages of the linking process.

As explored in sections 5.4.5 and 5.4.6, consequential and criterion-related validity are the two aspects of validity that received the least attention throughout the CEFR linking process. Lack of attention to consequences in linking to the CEFR or any external

benchmark might have several implications for institutions. For instance, analysis of consequential aspect of validity might lead to a change in the examination tasks, see for example the case of the City and Guilds Communicator project (O'Sullivan, 2009a), or curriculum changes. In the BUSEL study, the effectiveness of the writing syllabus was questioned considering the CEFR samples provided by the Council of Europe and the features of a B2 level writing as defined by the CEFR descriptors. As suggested above, organizations that are embarking upon a linking study should first of all question the suitability of the external criterion, the CEFR in this case, for their own context. For instance, detailed analysis of the CEFR was carried out to determine whether it would address the needs in Canada with regard to a common framework of reference for languages and a review of the current frameworks showed that the CEFR would be the best for use in Canada (Vandergrift, 2006). Similarly, the acceptable CEFR level for academic study in EAP contexts is commonly perceived to be B2 (PGMAC, 2012), the level COPE was linked at. Carrying out a linking study without considering and investigating its consequences would be meaningless because unless such a study has positive impact on the institution, for instance on classroom teaching and students in a school context, then institutions may need to consider the correctness of their decision to link their exams to the CEFR. In some cases the linking might only provide evidence supporting current practices in an institution. The linking process should at least lead to a better quality test.

In terms of criterion-related aspect of validity, those carrying out a linking study are advised to compare their exams with another exam that is already linked to the CEFR. However, not only does investigating the criterion-related aspects of validity involve comparison of an examination with an external test but also includes comparison with

the same form of the same examination in different administrations and alternative forms of the same examination in future administrations (Weir, 2005a). Therefore, the Manual could raise users' awareness in this respect and draw attention to possible ways of exploring consequential and criterion-related validity for their examinations.

Aside from the consequential criterion-related aspects of validity, consideration into the meaning of the results of the CEFR linking and what implications linking brings with it, or even what is learned from the process and how this knowledge can be used would be essential. Linking examinations to external standards have so much to offer to organizations that the Manual could perhaps suggest ways of how people can benefit from this process more fully. The implications of linking studies are further explored in the concluding chapter.

5.6 Summary

As indicated in the conclusion section of this chapter, the standardisation stage of the CEFR linking process seems to contribute most to the validity of an examination, as it requires an in-depth analysis of the examination under study and focuses on the greatest number of parameters in the Weir model. Moreover, even though the process helps question some aspects of validity, it neglects certain crucial aspects such as test taker characteristics. The results are discussed in detail in the concluding chapter (Chapter 7).

CHAPTER 6

PHASE 3 – REVIEW OF THE MANUAL APPROACH TO VALIDATION

6.1 Introduction

Phases 1 and 2 of this research, outlined in Chapters 5 and 6, investigated the implementation of the CEFR linking process in the BUSEL context based on Weir's validation framework, and collected data on the BUSEL approach to CEFR linking. Therefore, the conclusions drawn from these two phases reflect the practical side of the linking process as a unique adaptation of the methodology suggested by the Manual. For instance, the familiarisation stage, as opposed to the 3 hour session suggested in the Manual, lasted much longer and required seven different sessions. Some of the specification forms had to be filled in several times. Reading and writing standardisation sessions were held a number of times. The Manual was not followed to the letter because the group involved in the linking project was inexperienced in terms of CEFR and in order to make solid linkage claims some of the stages, such as the standardisation, were repeated until the group felt confident with the work they did in those stages. Phase 3 aimed to reflect on the validation approach implied in the Manual through the use of document analysis, a mini questionnaire and interviews.

In this chapter, the purpose of the review of the Manual approach to validation is given (Section 6.2) and then brief information on how the review was undertaken is presented (Section 6.3). The next section (6.4) presents the data analysis and findings of the critical review of the Manual. Finally, a summary of the chapter as well as conclusions are presented (6.5).

6.2 Purpose of the review of the Manual's approach to validation

As stated in the Manual, its primary aim is to help users link their examinations to the CEFR by developing, applying and reporting procedures in a quality manner, which involves describing the exam content and procedures for administration and test analysis; relating exam results to the CEFR; and providing evidence to support the quality of the procedures carried out throughout the linking process (Council of Europe, 2003: 1).

The Manual proposes a theoretical basis for a linking study through a set of suggested procedures, which may vary from one institution to another depending on how and/or whether the suggestions are implemented. For example, in the City and Guilds CEFR linking project (O'Sullivan, 2009a; 2009b; 2009c) a critical review stage was added to the linking process whereas in the case of the COPE linking, the stages reflected the ones suggested in the Manual. Phase 3 aims to review the approach to validation in the Manual in comparison with the BUSEL approach.

6.3 A critical review of the Manual approach to validation

The critical review of the Manual approach to validation was undertaken in three ways. Firstly, a document analysis of the Manual was carried to investigate how well the Manual captures aspects of validity suggested in Weir's framework. Secondly, a mini questionnaire was administered to explore the contributions of the linking process to the validity of the COPE examination, and thirdly, interviews were carried out to examine the impact of the institutional changes made to the Manual suggestions.

6.3.1 Document analysis

For the document analysis of the validation approach implied in the Manual, a chart with four columns was prepared. A section of the chart can be seen in Table 6.1 and the complete version of the chart can be found in Appendix 3M (See accompanying CD Folder 6).

Table 6.1 A section from the chart used for document analysis

WEIR'S VALIDATION FRAMEWORK		CEFR LINKING MANUAL	COMMENTS
TEST TAKER CHARACTERISTICS *directly connected to context validity (Weir, 2005:51)*	Physical / Physiological Characteristics **Does the test make suitable accommodations for candidates with special needs? (pg.53)*		
	Psychological Characteristics **In what ways does the test put the candidates at their ease? (pg.54)*		
	Experiential Characteristics **Are the candidates sufficiently familiar with what they have to do on the test? (pg.55)*		

- Column 1 – aspects of validity (e.g. context validity, criterion-related validity);
- Column 2 – validity parameters with key questions taken from Weir (2005);
- Column 3 – how each aspect of validity is tackled in the Manual as reviewed by the researcher;

- Column 4 – a comments section for people who checked the completed chart for confirmatory purposes.

(See section 3.7.3 for an overview of the procedures used to develop the chart for the document analysis)

This chart was produced for reading and writing separately, however, some of the sections were the same for both papers. For instance, in Weir's validation frameworks test taker characteristics do not change for reading and writing as the background knowledge and experience of students do not change from paper to paper. Similarly, parameters of context validity such as task setting, demands and administration procedures, are the same for both skills but how they are realized for writing and reading might differ. However, parameters for some other aspects of validity such as cognitive validity or scoring validity show variance over reading and writing. In such cases different charts were produced.

After the chart was prepared for all aspects of validity, the researcher analysed the Manual chapter by chapter, and filled in Column 3 by indicating at what stage of the linking process and how an aspect of validity is tackled in the Manual, if at all. Once the document analysis was carried out by the researcher the chart was given to two project members who reviewed the researcher's analysis of the Manual for confirmation purposes. They agreed or disagreed with the researcher as well as adding to or changing her findings.

Analysing the Manual in the way presented in Table 6.1 allowed for a clear presentation of what aspects of validity the Manual takes into consideration or implicitly encourages

users to consider and whether the conclusions from the review can also be drawn by people other than the researcher. This then enabled the researcher to see whether an implied approach existed in the Manual and its contributions to the validity of an examination regardless of the adaptations made to the procedures in the BUSEL context.

6.3.2 Questionnaire

The purpose of the questionnaire was to investigate the extent to which the Manual procedures contributed to the validity of the COPE, differentiating between the validity of COPE in general and as a result of the linking process. The questionnaire looked at all aspects of validity and asked the three respondents whether an aspect of validity was always considered for the COPE examination or whether the CEFR linking process contributed to that aspect of validity. The questionnaire can be found in Appendix 3N.

6.3.3 Interviews

The aim of the interview to investigate the extent to which the alterations made to the Manual linking procedures in the COPE project contributed to the linking process and the validity of the COPE examination. Three separate interviews were carried out with the respondents of the questionnaire and the data was analysed using codes emerging from the interview questions. The interview coding scheme can be found in Appendix 6A.

6.4 Data analysis and findings

In this section, findings regarding the critical review of the Manual are reported with reference to, first, the document analysis and the questionnaire (Section 6.4.1) as they

both focus on aspects of validity; then the interviews, which evaluate the institutional changes made to the Manual linking procedures. At times references are made to the people who reviewed the researcher's document analysis for confirmatory purposes, validator 1 (V1) and validator 2 (V2), as well as the people who were involved in answering the questionnaire and in the interviews (I1, I2, I3).

6.4.1 Findings of the document analysis and the questionnaire

6.4.1.1 Test taker characteristics

With respect to test taker characteristics, Weir draws attention to physical/physiological, psychological and experiential characteristics, referring to the work of O'Sullivan (2000).

(a) Physical/physiological characteristics

Weir proposes the question "does the test make suitable accommodations for candidates with special needs?" (2005a: 53). By special needs, he refers to candidates with certain disabilities such as hearing impairment or visual impairment. However, under physical/physiological characteristics, age and gender could also be listed; age in particular plays a significant role in the design of an examination. For instance, young learners' attention span must be taken into consideration at the design stage and examination aiming to test young learners.

The analysis of the Manual showed that test taker characteristics such as disability, gender or age are not considered in the CEFR linking process. The only place where age is mentioned is at the specification stage where the users are asked to define the target population of an examination in terms of age/grade (General Examination Form A1).

However, the real issue underlying such a specification, that is considering these characteristics with respect to how they influence test construction, is not addressed. The validators agreed with the researcher in this respect as summarized in Table 6.2.

Table 6.2 Summary of the review data for test taker characteristics

Test Taker Characteristics	Researcher	Validator 1	Validator 2
Physical/Physiological	Not tackled	Agree	Agree
Psychological	Not tackled	Agree	Limited
Experiential	Tackled	Agree	Agree

(b) Psychological characteristics

"In what ways does the test put the candidates at their ease?" (ibid: 54). Weir (2005a) believes that interest and motivation are important factors that influence performance in tests. This aspect of test design is not tackled in any way in the CEFR linking process. V2 added that the users of the Manual are asked to specify the communication themes students are expected to handle in form A10 for reading and form A14 for writing and the choice of themes is closely linked to interests or affective schemata. She also indicated that the cognitive style of the learners was discussed at the standardisation stage especially while defining the least able B2 candidate profile.

(c) Experiential characteristics

Experiential characteristics are related to candidates' familiarity with the examination they are going to sit. "Are the candidates sufficiently familiar with what they have to do on the test?" (ibid: 55). O'Sullivan (2006: 242) proposes that "familiarity with both task types and task performance conditions" have a positive effect on test taker performance.

The Manual deals with experiential characteristics at the specification stage of the CEFR linking process. It requires users to indicate the type of information that is published for candidates and teachers (General Examination Form A1). This seems to be sending users the message that test users should be informed about the test.

The questionnaire looked at test taker characteristics in general and the results, summarised in Table 6.3, corroborated the findings of the document analysis. The results showed that whereas several actions are taken for the COPE examination as explained in section 4.5, the CEFR linking process did not contribute to the test taker considerations of COPE.

Drawing the connection between the CEFR linking process and consideration of the test takers appears essential here. Test takers lie at the heart of any measurement event as exams are designed to measure their abilities or proficiency levels. Tests have to be suitable for their needs. Without a close scrutiny of the test taker profile and the due considerations of the characteristics indentified as being likely to impact test performance, valid tests cannot be designed. Therefore, the Manual needs to guide users to consider the test taker characteristics throughout a linking study so that appropriate tasks and activities can be chosen and an appropriate level can be set for the exam. Taking test taker characteristics into consideration throughout the linking process can help establish a close link between the needs of the target population and what the exam sets out to measure.

Table 6.3 Validity of COPE questionnaire results

VALIDITY ASPECT	READING			WRITING		
	ALWAYS	CEFR	NONE	ALWAYS	CEFR	NONE
TEST TAKER CHARACTERISTICS						
Physical/physiological	√ √ √ considered		√√√ evidence	√ √ √ considered		√√√ evidence
Psychological	√ √ √ considered		√√√ evidence	√ √ √ considered		√√√ evidence
Experiential	√ √ √ considered		√√√ evidence	√ √ √ considered		√√√ evidence
CONTEXT VALIDITY						
Fairness of test tasks	√ √ √ considered	√ considered	√√√ evidence	√ √ √ considered √ evidence	√ considered	√√ evidence
Fairness of test administration	√ √ √ considered		√√√ evidence	√ √ √ considered		√√√ evidence
COGNITIVE VALIDITY						
Cognitive processes interactionally authentic	√ √ √ considered	√ considered	√√√ evidence	√ √ √ considered	√ considered	√√√ evidence
SCORING VALIDITY						
Dependability of test scores	√ √ √ considered	√√√ evidence		√√√ evidence	√√√ evidence	
CONSEQUENTIAL VALIDITY						
Impact on stakeholders	√ √ √ considered		√√√ evidence	√ √ √ considered		√√√ evidence
CRITERION-RELATED VALIDITY						
External evidence	√ √ √ considered	√√√ evidence		√ √ √ considered		√√√ evidence

KEY: √ = 1 respondent

Always considered = areas BUSEL always takes into account and clearly documented

CEFR considered = areas the linking process helped consider

Always evidence = areas BUSEL has always collected evidence

CEFR evidence = areas the linking process helped gather evidence

None evidence = areas no evidence is available

6.4.1.2 Context validity

Weir (1993, 2005a) emphasizes the significance of context as a determinant of language ability and suggests that both performance conditions and operations of a test should be as close as possible to the real-life situation. Weir looks at context validity in three

aspects; task setting, administration setting, and task demands. The parameters underlying these aspects are dealt with in the following sections.

(a) Task setting

The parameters Weir looks at under task setting are summarized in Table 6.4.

Table 6.4 Task setting parameters

Context Validity Task Setting	Key questions
Rubric	Is the rubric accurate and accessible?
Purpose	Is the purpose of the test made unequivocally clear for the candidate? Is it an appropriate purpose?
Response format	Is there any evidence that the test response format is likely to affect the test performances?
Known criteria	Are the criteria to be used in the marking of the test explicit for the candidate and the markers?
Weighting	Are weightings for different test components adequately justified?
Order of items	Are the items and tasks in a test in a justifiable order?
Time constraints	Is the timing for each part of the test appropriate?

The users of the Manual are asked to indicate the type of information available (overall aim of the exam, marking/grading schemes, standardized samples showing pass level) to the candidates and teachers in terms of *purpose* and the *known criteria*, only on specification form A1, which is designed to offer only an overview of a test. However, the purpose of each paper, for instance the reading paper, is not questioned separately. The *weighting* and *time constraints* are also covered in the specification forms in a similarly limited manner by simply indicating whether the candidates are informed of these or not. Similarly, regarding the *response format,* the users are asked to specify the response format in the specification form A1. None of these parameters are questioned in terms of their *adequacy*, which is the main focus in Weir's validation frameworks for reading and writing (as well, of course, as those for speaking and listening). Questions

such as whether the purpose of the test is appropriate for the target situation or whether the weightings of different sections of the test are well thought through are key questions in test construction and validation. V1, agreeing with this comment, further explains that the questions in the specification forms do not consider if the test tasks reflect real life tasks in terms of creating appropriate conditions and operations and adds "if the testing board cannot identify or find ways to identify those, they cannot judge a candidate's ability to function using the target language".

It should also be indicated that issues related to *rubric* and *order of items* are not tackled at all. However, as indicated in Table 6.5, V2 pointed out that CEFR specification form A1 asks questions about whether test tasks and sample papers are available to students and teachers. Simply the fact that students get a chance to look at reading or writing tasks and sample papers may solve the problems of *rubric, weighting, order of items* and *time constraints* in that students get familiar with them and can clarify ambiguity related to these issues, though without specific instructions or recommendations to use the sample papers in this way the developer cannot be in any way certain that the intended population of candidates will equally benefit. However, the issue of whether the weighting, for instance, or the order of items is justifiable is still not addressed.

Table 6.5 Summary of the review data for context validity task setting parameters

Parameters of context validity	Researcher	Validator 1	Validator 2
Rubric	Not tackled	Agree	Limited
Purpose	Limited	Agree	Agree
Response format	Limited	Agree with further explanations	Agree
Known criteria	Limited	Agree	Agree
Weighting	Limited	Agree	Agree
Order of items	Not tackled	Agree	Limited
Time constraints	Limited	Agree	Agree

(b) Administration setting

The parameters of administration setting are related to physical conditions, uniformity of administration and security as presented in Table 6.6.

Table 6.6 Context validity administration setting parameters

Context Validity Administration Setting	Key questions
Physical conditions	Were the physical conditions of the test administration satisfactory?
Uniformity of administration	Was the test administered in the same manner across sites?
Security	Was the test secure?

Issues regarding the administration of tests are not tackled in any way in the CEFR linking process as agreed by the validators and indicated in Table 6.7. However, the way a test is administered might have a huge impact on the reliability of the test as well as the performance of test takers. For instance, if a test is administered in a room where there is not enough light or where it is too cold, the performance of the test takers might drop. Another example is related to the uniformity of administration. Weir (2005a) talks about *uniformity of administration* across sites, however uniformity across occasions is also as important. If slightly more time than specified is given to the test takers in one

exam room or in one occasion, this raises questions regarding reliability and fairness of that test.

Even though there is no mention of administration systems in the Manual, it clearly states that one of its aims is to raise awareness "among developers of quality language examinations" (Council of Europe, 2003: 29) particularly at the specification stage. Therefore, in order to fulfil this aim, the Manual could provide more guidance on administration of exams.

Table 6.7 Summary of the review data for context validity administration setting parameters

Parameters of context validity	Researcher	Validator 1	Validator 2
Physical conditions	Not tackled	Agree	Agree
Uniformity of administration	Not tackled	Agree	Agree
Security	Not tackled	Agree	Agree

(c) Task demands

The first parameter of task demand as presented in Table 6.8 is *discourse mode,* which focus on categories of genre, rhetorical task and patterns of exposition. Specification forms A9 to A14, dealing with the detailed description of sub-tests, require that the developer indicates the discourse mode. Discourse mode is referred to as 'domains' in the forms and the users of the Manual are asked to refer to the CEFR, where possible text types the learners might come across in certain domains (broadly classified as public, personal, educational and occupational) are listed. However, genres, rhetorical tasks and patterns of exposition are not specified explicitly. It should also be noted that the appropriateness of the discourse mode is not questioned in the forms, which again is

what Weir encourages testers to explore. With respect to this, V1 comments that "the degree of authenticity and the organization of content are of great importance when choosing a reading text and should therefore be carefully defined". She continues to comment that the relationship between text type and response format is not questioned in the Manual either.

Table 6.8 Task demands parameters

Context Validity Task Demands	Key questions
Discourse mode	Is the discourse mode appropriate for the skills or strategies being tested?
Channel	Is the channel appropriate for the target situation requirements of the students being tested?
Text length	Is the test length appropriate for the target situation requirements of the students being tested?
Nature of information	Is the type of information appropriate for the target situation requirements of the students being tested?
Content knowledge	Is the topic content appropriate for the target situation requirements of the students being tested?
Lexical knowledge	Are the lexical items in the test both in input text and required as output appropriate for the level of the candidates?
Grammatical knowledge	Are the grammatical items in the test both in input and required as output appropriate for the level of candidates?
Functional knowledge	Are the functions in the test both in input and required as output appropriate for the level of the candidate?

The issue of *channel* is not tackled in any way by the Manual and V1 gives a further explanation as seen in Table 6.9, indicating that it is not quite relevant to the reading paper unless the type of information in the texts is of concern. However, besides the type of information, *channel* also refers to the layout or format of the reading texts, the writing prompts or the paper as a whole. If the test consists of several short texts in the

form of authentic advertisements, then the channel may have an impact on student performance with respect to reading. For instance, in her PhD study, dos Santos (2005) found that manipulation of the physical presentation of the text could significantly impact on test performance. Another example would be asking students to analyse and synthesize a graph and write a report about it. Those candidates who are not good with such visuals may be at a disadvantage.

Table 6.9 Summary of the review data for context validity task-demand parameters

Parameters of context validity	Researcher	Validator 1	Validator 2
Discourse mode	Limited	Agree with further explanations	Agree
Channel	Not tackled	Agree	Agree
Text length	Limited	Agree with further explanations	Agree
Nature of information	Not tackled	Agree	Agree
Content knowledge	Not tackled	Agree	Agree
Lexical knowledge	Limited	Agree	Agree
Grammatical knowledge	Limited	Agree	Agree
Functional knowledge	Limited	Agree	Agree

As for *text length*, this parameter is associated with proficiency levels in the CEFR scales. For instance, as a learner goes up the CEFR ladder, s/he becomes proficient enough to cope with relatively longer texts. While filling in the specification forms, the users of the Manual are asked to specify the text length. During the standardisation stage, the judges are required to consider text length when assigning levels to items attached to a text as it is part of the CEFR descriptors. However, appropriacy of text length in terms of target situation requirements is not addressed in the Manual. V1 also suggests that appropriacy of text length is not defined in the CEFR descriptors. The

term ‘lengthy text’ as described in the C1 Overall Reading Comprehension scale can differ from one person to another and one context to another. The same stands true for writing ‘short simple essays’ as in the B1 Essays and Reports scale. Lengthy for *what* and lengthy for *who* are the questions that need to be addressed in relation to ‘appropriacy’ questions.

In Weir’s framework (2005a:), *nature of information* is related to the information being abstract or concrete, which is covered under the concept of ‘complexity’ in the CEFR. At the standardisation stage, the judges are required to consider the complexity of the texts (C1) the learners are asked to tackle through the reading tasks or whether they can produce complex texts (C2) in a writing test. However, appropriateness of the nature of information in the text in terms of target situation, as validators agree, is not questioned in the Manual.

Content knowledge is part of the ‘communication themes’ in forms A9 to A14 where the topics learners need to handle are listed. However, again as the validators agreed, the appropriateness of these is not questioned.

Finally, in terms of *lexical, structural and functional* requirements of the test, these parameters are dealt with in forms A19 for reading and A21 for writing, involving aspects of language competence in reception and production. These are linguistic (lexical and grammatical), socio-linguistic, pragmatic (macro and micro functional competences) and strategic competences. With respect to the appropriateness of these for the level of the learners, the relevant specification forms ask the users to analyse these competences and decide at which CEFR level the test can be situated based on

these features of the test. In fact, the forms do just the opposite of what Weir suggests. In the forms, neither the level intended by the test developers nor whether it is reflected in the test is questioned. This raises issues of construct validity as the forms merely focus on the level of a test in terms of the CEFR, not whether there is a (mis)match between the intended level and the level the test is at. Lexical, structural and functional requirements of the test are examined once again at the standardisation stage of the linking process when the judges are asked to assign CEFR levels to reading items or written performances. These judgments entail a close scrutiny of the lexical, structural and functional requirements of the test both at the input and output level. In other words, the requirements of tests needs to be examined separately for the texts used as input and the texts that test takers are required to produce in writing.

In terms of the questionnaire results, only one of the respondents thought that task demands were considered for the COPE examination during the CEFR linking process, which might suggest that there is little consideration of context validity parameters in the CEFR linking process (Table 6.3). The other two respondents may not have considered aspects of context validity as covered through the linking process because of the areas that were either not tackled in the Manual or only to a certain degree, as shown by the results of the document analysis.

One might argue that thinking within the parameters of the context aspect of validity at certain stages of the CEFR linking process does not necessarily validate an examination in this respect. However, throughout these stages, if the discussions held are recorded and documented in some way, they could be considered as a piece of evidence regarding the context aspect of validity. The Manual suggests that each stage should be

documented as to what has been done and how with reasons, however, it does not go into details as to what is meant by the term *documentation*. In the COPE project, as well as the session notes or relevant forms, the discussion were also documented which could be analysed and provided as evidence towards the validity of the COPE examination.

6.4.1.3 Cognitive validity

Cognitive validity entails understanding the cognitive processing involved in the performance of a test task (Weir, 2005a). Cognitive validity is analysed in two aspects: executive processes (Table 6.10) or cognitive processing as indicated in the latest version of the model (2009) and executive resources (Table 6.11) or cognitive load (2009). It also has different parameters for cognitive processing, as specified in Table 6.10, for reading and writing that come from the very nature of receptive and productive skills. Executive resources on the other hand are the same for both skills. The summary of the data for both executive processes and resources is given in Table 6.12.

Table 6.10 Cognitive validity executive processes (cognitive processing)

Cognitive Validity Executive Processes	
Reading	Writing
Goal setting	Goal setting
Visual recognition	Topic/genre modifying
Pattern synthesizer	Generating
	Organizing
	Translating

Throughout the CEFR linking process, issues related to *executive processes* are mainly dealt with in Chapter 6 (Empirical Validation) of the Manual. In the internal validation section, it is recommended that qualitative methods such as reflection, analysis of samples or feedback methods are used to investigate these (Council of Europe, 2003).

Guidance on qualitative analysis methods is given in the Reference Supplement to the Manual in Section D (Banerjee, J. 2004). In addition, at the specification stage of the linking process, the users are asked to specify what kind of strategic competences the test takers are expected to be able to handle in terms of *monitoring* for reading in form A19 (Aspects of language competence for reception) and *generating, organizing, translating and monitoring* for writing in form A21 (Aspects for language competence for production). These include planning, execution, evaluation and repair.

Table 6.11 Cognitive validity executive resources (Cognitive load)

Cognitive Validity Executive Resources (Cognitive load)	
Language knowledge	Grammatical (lexis, syntax)
	Textual
	Functional (pragmatic)
	Sociolinguistic
Content knowledge	Internal
	External

When it comes to the *cognitive load*, similar to *cognitive processing*, such issues are mainly dealt with in Chapter 6 of the Manual. There is a recommendation that qualitative analysis methods are used to explore these issues (ibid). In the specification stage also, the users are asked to specify what kind of language knowledge, including linguistic, socio-linguistic and pragmatic competences (discourse and functional) students are expected to be able to handle in form A19 (Aspects of language competence for reception) and A21 (Aspects of language competence for production). In addition, at the standardisation stage judges are required to consider the executive resources test takers need to possess to be able carry out the exam tasks. Issues related to *content knowledge* are not tackled in the CEFR linking process in any way.

In terms of cognitive validity, the Manual draws users' attention to cognitive processing and load; and encourages them to collect evidence regarding cognitive validity of the exam in question.

Table 6.12 Summary of the review data for cognitive validity

Parameters of cognitive validity	Researcher	Validator 1	Validator 2
Executive processes - Reading			
Goal setting	Tackled	Agree	Agree
Visual recognition	Tackled	Agree	Agree
Pattern synthesizer	Tackled	Agree	Agree
Executive processes – Writing			
Goal setting	Tackled	Agree	Agree
Topic/genre modifying	Tackled	Agree	Agree
Generating	Tackled	Agree	Agree
Organizing	Tackled	Agree	Agree
Translating	Tackled	Agree	Agree
Executive resources			
Language knowledge	Tackled	Agree	Agree
Content knowledge	Not tackled	Agree	Agree

Again, only one of the respondents to the questionnaire indicated that parameters of cognitive validity were considered for the COPE examination during the linking process (Table 6.3). This might be because the other respondents did not think that the CEFR process in fact helped improve the cognitive validity aspect of COPE or collect data in this regard, as shown by the document analysis. However, the same argument presented for context validity is also relevant here. Documented discussions of the CEFR sessions may also serve as a piece of evidence towards the cognitive aspect of validity for COPE. The Manual is already providing methods to collect qualitative evidence on the exam itself and the linking process through Section D of the Reference Supplement (Banarjee, 2004), which could help documenting discussions in a systematic way.

6.4.1.4 Scoring validity

Scoring validity includes the concepts of both making a test and its scoring more reliable (Weir, 2005). Making a test more reliable involves carrying out item analysis and internal consistency analysis to make sure the items work well at an item level and the test functions well at a more global level. It has different parameters for reading and writing due to the differing nature of the tasks used to measure these skills (Table 6.13). Table 6.14 summarizes the data for reading and writing.

Table 6.13 Scoring validity parameters

Scoring Validity	
Reading	Writing
Item analysis	Criteria/rating scale
Internal consistency	Rating procedures (training, standardisation, conditions)
Error of measurement	Raters
Marker reliability	Grading and awarding

For reading, *scoring validity* is undertaken at the empirical validation stage of the CEFR linking process as part of the internal validation where data on reading items, internal consistency and error of measurement need to be gathered either using CTT or IRT. As for marker reliability, the suggested methods are reflection and feedback, which represent a qualitative analysis method. Another method suggested is carrying out a generalisibility study to determine the optimal number of markers needed and the most suitable people for marking so as to improve the reliability of the exam. It is also recommended that an item bank should be set up in order to ensure parallel tests. The Reference Supplement to the Manual also offers guidance on CTT, Generalisibility, Factor Analysis and IRT (Verhelst, N. 2004a; 2004b; 2004c; 2004d).

Whether scoring validity is purely in the domain of internal validity, as suggested by the Manual, is another issue and will be discussed in the conclusion of this chapter.

Table 6.14 Summary of the review data for scoring validity

Parameters of Scoring validity	Researcher	Validator 1	Validator 2
Reading			
Item analysis	Tackled	Agree	Agree
Internal consistency	Tackled	Agree	Agree
Error of measurement	Tackled	Agree	Agree
Marker reliability	Tackled	Agree	Agree
Writing			
Criteria/rating scale	Tackled	Agree	Agree
Rating procedures	Tackled	Agree	Agree
Raters	Tackled	Agree	Agree
Grading and awarding	Tackled	Agree	Agree

The results of the questionnaire supported the findings of the document analysis. All the respondents of the questionnaire indicated that the CEFR linking process helped collect evidence towards the scoring aspect of validity for the COPE examination.

6.4.1.5 Consequential validity

Consequential validity involves "the adequacy and appropriateness of interpretations and actions based on test scores" (Weir, 2005a: 210). Its parameters are presented in Table 6.15; it requires investigating whether the test is biased towards a certain group (differential validity), what the impact of the test is on classroom practices and workplaces; and finally what effect it has on test takers as members of a society.

Table 6.15 Consequential validity parameters

Consequential Validity
Differential validity
Washback in classroom or workplace
Effect on individual within society

None of these aspects of consequential validity, as presented in Table 6.16, is tackled in the CEFR linking process by the Manual. However, in the qualitative analysis methods of the Reference Supplement to the Manual, an impact study is given to demonstrate the usefulness of certain qualitative research tools. Being only an example, this may not be effective in communicating the message that impact studies are of great importance to test validity. Whether a test has positive or negative impact on teaching materials, classroom activities, learning and test takers should all be investigated because, as Messick (1989) points out, empirical evidence should be gathered to support the adequacy and appropriateness of interpretations and actions based on test scores.

Table 6.16 Summary of the review data for consequential validity

Parameters of Consequential validity	Researcher	Validator 1	Validator 2
Differential validity	Not tackled	Agree	Agree
Washback	Not tackled	Agree	Agree
Effect on individual	Not tackled	Agree	Agree

In terms of the questionnaire results, as indicated by all the respondents, the impact of the COPE examination on the stakeholder, though always taken into consideration in BUSEL, was not analysed as part the CEFR linking process (Table 6.3).

6.4.1.6 Criterion-related validity

Criterion-related validity entails demonstrating a relationship between a test and an external criterion measuring the same ability (Weir, 2005a). External criterion has a broad meaning in this context. It could be a different version of the same test, the same test administered on different occasions, other tests or measures such as the CEFR which is a set of reference levels, or future performance as listed in Table 6.17.

Table 6.17 Criterion-related validity parameters

Criterion - related Validity
Comparison with different versions of the same test
Comparison with the same test administered on different occasions
Comparison with other tests/measures
Comparison with future performance

The Manual encourages setting up an item bank and the use of IRT for internal validation as IRT allows for comparison with different versions of the same test. It also encourages users to compare the test in question with an external test that is already properly linked to the CEFR or to teacher judgments using the CEFR scales as a basis in external validation, though whether the use of teachers' judgments is valuable as a validation tool is up for discussion (to be addressed in the conclusion of this chapter). V2 clarifies that teacher judgments are suggested by the Manual but that calibrated tests with different purposes, i.e. academic, are not available as external criterion. The other aspects of criterion-related validity are not tackled in the CEFR linking process as indicated in Table 6.18.

Table 6.18 Summary of the review data for criterion-related validity

Parameters of Criterion-related validity	Researcher	Validator 1	Validator 2
Different versions	Tackled	Agree	Agree
Different occasions	Not tackled	Agree	Agree
Other tests	Tackled	Agree	Agree
Future performance	Not tackled	Agree	Agree with a further explanation

Criterion-related aspect of validity is another area all the respondents of the questionnaire agree that the CEFR helped gather validity evidence for the COPE examination (Table 6.3).

6.4.2 The results of the interviews

The aim of the interviews was to investigate the extent to which the alterations made to the Manual linking procedures in the COPE project contributed to the linking process and the validity of the COPE examination. The tables provided in the subsequent sections present the themes and the codes related to them, and indicate which interviewee(s) raised the same codes. As the interview questions focused on each stage of the CEFR linking process, the results of the interviews are presented following these stages. At times direct quotations from the interviews are used to support a point and in such cases the interviewee codes (I1, I2, or I3) are indicated and reference to the relevant lines in the transcriptions are provided (See Appendix 6A for the interview coding scheme).

6.4.2.1 Familiarisation stage

The adaptations made to the Manual suggestions for the familiarisation stage mainly involved extending the time framework, as opposed to a 3-hour session, thus enabling

the project members to have a firm background to the CEFR and further analyse the CEFR scales and descriptors. Another significant addition to the recommended process was the gathering of data from the tasks undertaken by the participants, such as rank ordering scales or quizzes, which were analysed using IRT and the results shared with the participants.

As summarized in Table 6.19, the interviewees found the decision to extend the familiarisation stage of the process and carrying out statistical analysis at this stage beneficial and essential to move forward as not only did it cause the project to be seen as an indication of the commitment of BUSEL to improving the quality of the COPE but also contributed to the validity of the stage. Knowing the background to the CEFR and having more opportunities to work with the scales helped participants better understand how to apply them. I1 indicated that "the validity of what we were trying to do was in direct relationship to how we interpreted the CEFR benchmarks so that was an important part of the process" (I1: 31-31).

Table 6.19 Phase 3 Interview - Familiarisation

Themes	Descriptions	I1	I2	I3
Validity of the CEFR linking process	Using the CEFR	√	√	√
	Forming connections	√	√	
	Confidence		√	√
	Seriousness		√	√
Validity of COPE	Using the CEFR			√

The familiarisation stage may not have had a direct impact on the validity of the COPE exam, scoring validity in particular, but as I3 emphasized the importance, thus the impact, of the familiarisation stage at later stages of the project by saying, "if I couldn't

understand what the CEFR was thoroughly, it would probably affect the judgments that I made" later on regarding the COPE exam (I3: 324-325).

6.4.2.2 Specification stage

At the specification stage, rather than the people responsible for the COPE exam completing the specification forms as suggested in the Manual, the whole CEFR linking project group filled them in jointly. In addition, some of the forms were filled in during a formal session whereas some others had to be completed outside a session. Completion of the forms outside a session might have led the participants interpret the questions in different ways thus resulting in different outcomes since they did not have a chance to discuss the questions with one another and reach a common understanding. This issue was further discussed in section 4.3.2.3. In the case of individual completion of the forms, the data were collated and the forms were sent back to the group members several times with feedback until consensus was reached.

Respondent I3 felt that filling in the forms in a formal session helped her do a better job as she had had a chance to clarify the sections or terminology she could not understand. This contributed to the validity of the process and the COPE exam because "If we hadn't filled in the forms properly, it would again be a problem while we were allocating levels and it would affect the validity negatively" (I3: 359-360). The interviewees felt, as summarized in Table 6.20, that completing the specification forms as a group and revising them several times made the process a more valid one. The BUSEL approach "contributed to the validity of the COPE exam by making us question things, clarify things, and add slightly to the [test] specifications" (I2: 199-201).

Table 6.20 Phase 3 Interview - Specification

Themes	Descriptions	I1	I2	I3
Validity of the CEFR linking process	Group effort	√	√	√
	Better understanding of COPE	√	√	
Validity of COPE	Assigning levels			√

6.4.2.3 Standardisation stage

The Manual provides guidance in the Reference Supplement as to the standard setting methods and how they might be applied. Having presented a number of options, the Manual writers finally recommend the use of the Basket approach, which requires judges to answer the question "At what CEFR can a learner answer this item correctly?" Apart from basic descriptive statistics required to calculate cut scores, no additional statistical data were suggested to be collected in the Manual. In the COPE linking study, more than one standard setting method was used and statistics related to inter-judge reliability including MFR were employed. The standardisation sessions were also carried out a number of times until the group felt confident with the cut scores established.

Regarding carrying out both the writing and reading standardisation sessions a number of times to ensure a high level of judge agreement, all three interviewees thought that judges gained experience going through the process several times and felt more confident about their judgments. This had an impact on the validity of both the standardisation stage and the COPE as confident claims could be made regarding the cut scores established. As presented in Table 6.21, the use of statistics such as MFR had a significant role in this as they enabled the participants to take the process seriously (I2: 246) and allowed for the analysis of the standard setting methods in order to establish viable cut scores (I1: 71-72).

Table 6.21 Phase 3 Interview - Standardisation

Themes	Descriptions	I1	I2	I3
Validity of the CEFR linking process	Cut scores	√	√	
	Advanced statistics	√	√	√
	Confidence	√	√	√
Validity of COPE	Cut scores	√	√	
	Advanced statistics	√	√	√
	Confidence	√	√	√

In terms of the use of more than one standard setting method, both I1 and I2 thought that comparison of methods allowed for more dependable cut scores to be established.

6.4.2.4 Empirical validation stage

For the empirical validation stage of the CEFR linking process, the Manual provides information and guidance regarding a number of analyses for internal validity and external validity, both through the Manual itself and the Reference Supplement. CTT, IRT, factor analysis, generalisability and qualitative analysis of the exam are among the suggestions for internal validity and linking to an external criterion such as another CEFR calibrated test or teacher judgments were recommended for external validation. In the BUSEL approach to validation, although a broader validity perspective was embraced, some of the suggestions in the Manual such as CTT, IRT and linking to an external criterion were carried out, some others could not be done due to limitations regarding time and resources. However, sending written scripts to external people to be marked and making teacher judgments and linking to external exams a continuous and integral part of the COPE analysis, were the additions made to the Manual suggestions. These are discussed briefly below.

Sending written scripts to external people to assign levels (which are based on the CEFR descriptors), as indicated by two interviewees, allowed for more objective

decisions regarding the COPE writing cut score. Using teacher judgments and other exams as external criteria were useful not only for the linking process but also for the validity of the COPE examination as they contributed to the criterion-related validity argument for the examination. Teacher judgment "has become a part of the institutional culture" (I2: 283) whereby teachers at BUSEL are encouraged to assess their students in relation to the CEFR prior to every COPE administration. As summarized in Table 6.22, two of the interviewees also stated that teachers have become more aware of the expectations inherent in COPE and those implied by the B2 level. Ongoing linking of examinations such as FCE, CAE or the Communicator exam to COPE contributes to the growing criterion-related validity evidence for the COPE examination.

Table 6.22 Phase 3 Interview – Empirical Validation

Themes	Descriptions	I1	I2	I3
Validity of the CEFR linking process	Institutional bias	√	√	
	Teacher judgments		√	√
	External exams		√	√
Validity of COPE	External people		√	
	Expectations		√	√
	Evidence		√	√

6.5 Summary and conclusions

This chapter has looked at the CEFR linking process as suggested in the Manual and as implemented in the COPE linking project in relation to Weir's validation framework. Firstly, each aspect of validity for reading and writing were investigated to see how and how well the Manual captures them through a document analysis; the contributions of the linking process to the COPE examination were identified through a questionnaire. Secondly, the alterations made to the CEFR were evaluated through interviews.

Phase 3 of the research suggested that the Manual procedures had only a limited contribution to offer in terms of validation. In terms of test taker considerations, with the exception of experiential characteristics, test taker characteristics were not considered, which was also borne out in Phase 1 Section 4.3.5 and Phase 2 Section 5.4.1, where it was concluded that test taker characteristics do not seem to have a place in the linking process.

Regarding the context aspect of validity, the document analysis and the questionnaire showed that the Manual procedures had a limited focus on context parameters. In Phase 1, the issue of context seemed to come up in the familiarisation stage (Section 4.2), specification (Section 4.3) and standardisation (Section 4.4) stages of the linking process. The limited contribution of the process to the context aspect of validity was also confirmed in Phase 2 Section 5.4.2 where standardisation seemed to be the stage task demands were most considered.

The results of the document analysis showed that the cognitive aspect of validity was tackled indirectly whereas the questionnaire results indicated that the cognitive aspect was not covered through the linking process. The indirect nature of the tackling of cognitive parameters was brought out in Phase 1 sections 4.2, 4.3, 4.4; and Phase 2 section 5.4.3, where it was shown that apart from empirical validation, parameters of the cognitive aspect of validity are examined both for the reading and writing papers of COPE at all stages of the linking process. Especially in specification and standardisation stages, these parameters were highly important in defining the level of the papers.

In terms of scoring and criterion-related aspects of validity, analysis of the questionnaire results showed that the linking process as suggested by the Manual contributed to these aspects of validity for COPE and the Manual analysis revealed that whereas scoring validity is fully tackled in the process, the criterion-related aspect of validity is limited. Regarding consequential validity, both the document analysis and the questionnaire indicated that this is an aspect that is not considered in the linking process. Findings regarding scoring, consequential and criterion-related aspects of validity were also borne out in Phases 1 and 2 of the research.

The interviews showed that the alterations made to the Manual linking procedures were perceived positively and contributed to the validity of the linking process and the validity of the COPE examination itself. Extending the familiarisation stage contributed to the linking process by helping project members better understand the CEFR and use the scales, which had an effect on the success of the subsequent stages and contributed to the scoring validity of both the linking process and the COPE. Filling in the specification forms as a group, including people who were not constructors of the COPE examination, in addition to working on some of the forms for an extended period of time made the stage more reliable and contributed to the COPE by further developing the test specifications, strengthening context and particularly cognitive aspects of validity. At the standardisation stage a number of changes added to the accuracy of the decisions made, these included:

- The use of more than one standard setting method helped confirm the cut scores set and helped ensure that those participating in this stage of the project were more confident of the final cut score than they might have been if a single method had been used.

- Holding the standard setting sessions several times to reach high level of judge agreement enabled the participants to understand the B2 level better and make more confident judgments about the exam.
- Carrying out advanced analysis on the standard setting data at the standardisation stage helped participants become standard and set viable cut scores.

These alterations enhanced the scoring validity of the standard setting and in return the scoring validity of COPE.

The alterations made at the empirical validation stage such as involving external people to judge written scripts, making teacher judgments, and linking to external exams provided the institution with criterion-related validity evidence for COPE.

While the Manual specifies what its aims are, it also clearly indicates what it does not set out to do:

- "*It provides a guide specifically focused on procedures involved in the validation of a claim that a certain examination or test is linked to the CEF.*
- *It does not provide a general guide on how to construct good language tests or examinations. There are several useful guides that do this and they should be consulted. Relating examinations to the CEF makes sense only if the examinations are of good quality.*
- *It does not prescribe any single approach to constructing language tests or examinations. While the CEF espouses an action-oriented approach to language learning and use, being comprehensive, it accepts that different examinations reflect different goals (constructs). Before embarking on relating examinations*

to the CEF, it is the prior responsibility of the examination providers to demonstrate the validity of their examination by showing that it assesses the constructs intended." (Council of Europe, 2003: 1).

If the main purpose of the Manual is to offer procedures as to how an examination can be linked to the CEFR and how validation of the linkage claim can be made, the question as to how the Manual can distance itself from any validation theory arises. Inevitably, CEFR linking still involves 'validation' no matter what the focus is and a theory is vital to ensure systematic gathering of the sort of evidence required to create a convincing validation argument and linkage claim.

The bullet points listed above that specify what aims the Manual has and does not have, contradicts to a certain degree what is actually suggested in Chapter 6 of the same Manual. Not only does Chapter 6 deal with validation of a linkage claim, but it also offers guidelines on how evidence on the internal validity of the examination being studied can be collected. Whereas it may be understandable why the Manual does not offer a validation theory to work with (the authors may not want to show bias towards or against any current model or may be taking into consideration the possibility of a more robust model emerging in the future), it is amongst the most profound responsibilities of the Manual to tell its users that a validation theory is vital to any project encompassing a test or an examination and recommend the use of an appropriate theory for validation purposes, be it solely for validating the linkage claims made. This is a significant shortcoming of the Manual, especially when most aspects of validity are already tackled through the CEFR linking process, as has been presented in this chapter and in Chapter 5.

Certain issues need special attention in terms of the link between validation and the CEFR linking process suggested by the Manual. One of these issues is related to context validity. The CEFR is claimed to be context-free (Council of Europe, 2001), however, considering context in the sense of context validity, an underlying construct of the context implicitly set through the CEFR exists. When examinations are aligned to the CEFR, the standards, and thus the context parameters set through the CEFR are embraced. For instance, the CEFR contains detailed specifications regarding content knowledge, which is a task-demand parameter under context validity. A2 involves a range of topics of most immediate relevance (e.g. very basic personal information, family information, shopping); B1 involves work, leisure, school; and B2 includes topics in the field of specialization. Another example is the nature of information and regarding this, abstract topics come into play at B2 level. Not all aspects of context validity are relevant to this discussion, however, attention could be drawn to this issue and the users of the Manual could consider how many of the parameters set through the CEFR are relevant and appropriate to their own context (e.g. the academic context), through the specification forms and standard setting as suggested in Chapter 5. The same discussion is true for cognitive validity since most of the parameters of both aspects of validity, cognitive load in particular, are in common. Moreover, the test taker, an inseparable part of test construct, could also be taken into consideration bearing in mind that it is the test taker, context, cognitive and scoring aspects of a test and the interaction among them that shapes the construct of a test (O'Sullivan & Weir, 2011).

Another issue that requires special attention concerns scoring validity. Certain aspects of scoring, such as reliability of marking or rater analysis, are dealt with in the empirical validation stage of the CEFR linking process. However, the place of scoring validity in

this process is broader than just empirical validation. At the familiarisation stage, rank ordering tasks can be analysed in order to accumulate data as to the performance of the judges, for instance, which will then feed into evidence supporting the quality of the linkage process. Similarly, at the standardisation stage, the reliability of the judgments made while setting cut scores must be analysed. These are all within the domain of scoring validity and crucial to the justifiability of the cut scores established and the CEFR linkage.

In terms of empirical validation, the Manual does not make any differentiation between productive and receptive skills. Therefore, as some of the tools suggested such as CTT or IRT only work for receptive skills, it is difficult to detect to what extent criterion-related validity such as comparison with different versions of the same test for writing is addressed in the Manual. Using teacher judgments, an examinee-centred standard setting method, is an appropriate tool for empirical validation, criterion-related validity in particular. Although it is not presented as a validation tool in the Manual, it can still be argued that the use of multiple methods is a good approach in order to reach a more reliable cut score than relying on a single method. In this case teacher judgments have an indirect effect on the validity of the examination in question through validation of the proposed cut score in relation to the CEFR.

Different aspects of validity gain importance at different stages of the CEFR linking process. For instance, while the specification stage mostly deals with parameters of context validity, investigating cognitive validity comes into play at the empirical validation stage. In addition, scoring validity is spread through almost all stages although this is not stated explicitly. However, the connections between these various

aspects of validity are lacking in the Manual. Having reviewed some of the issues regarding specific aspects of validity, it appears essential to highlight that the Manual lacks a cohesive theoretical structure that connects the various aspects of the Manual approach in terms of validation. This issue has been raised by O'Sullivan (2009a, 2009b, 2009c) in a series of City and Guilds linking study reports, which is discussed in the Concluding chapter of this research. Various aspects of validity are tackled in the Manual; however, they are not presented as a cohesive whole, as current validity theory suggests (Messick 1996; Mislevy et al., 2003: 2004: Weir, 2005a). Therefore, different elements of validity evidence may be collected but they are not systematic enough to build an integrated validation argument. Moreover, the role of standard setting or linking in a validation argument is not clarified at all by the Manual (This issue is also discussed in the concluding chapter). This results from the lack of an underlying theory of validation which diminishes the impact of the Manual and any linking studies which rigidly follow the procedures suggested there. It is for these reasons that the COPE project, seeing the role of linking and standard setting in validation, adopted a validation theory, that of Weir (2005a), right from the beginning of the project as presented in section 5.5.

CHAPTER 7

CONCLUSION

7.1 Introduction

This final chapter draws conclusions based on the case study presented in the preceding chapters. It first summarizes the research and its findings, and then presents the limitations of the study. This is followed by a discussion of the implications of this study for proficiency exams, institutions with a series or 'suite' of exams, the CEFR linking Manual and validation frameworks. Next, the contributions of this study to the field of assessment are presented. Finally, some concluding remarks are offered.

7.2 Summary of the research

7.2.1 Background

With the recent growing interest in the CEFR, the number of institutions linking their examinations to the CEFR has also increased, demonstrated by publications such as Figueras and Nijons (Eds) 2009 and Martyniuk (Ed.) 2010, both of which include several linking studies from across the world. As indicated in Sections 2.5 and 2.6, with the exception of a limited number of studies (e.g. O'Sullivan, 2009a, 2009b, 2009c), those linking their examinations to the CEFR did not investigate the contributions of the process to their tests with respect to validation. Another lack in the literature is the impact of such studies on pre-determined standards set through those tests. Therefore, this study aimed to investigate the role of linking in the validity argument as well as in establishing the level of an examination.

7.2.2 Research findings by question

In this section, the research findings are summarized for each research question with reference to the three phases of the study. In order to facilitate the investigation of the research questions, they were broken into sub-questions and findings for each are as follows:

7.2.2.1 Research Question 1

Does linking an examination to the CEFR provide a comprehensive validation argument?

1a. To what extent are the test taker characteristics taken into consideration during the linking process?

Test validity involves making accurate and meaningful interpretations of test scores, hence the inferences drawn about test takers (Carmines & Zeller, 1979; Messick, 1995; 1996; Weir, 2005a). In the case of a CEFR linking study, decisions are made about the language proficiency levels of test takers in relation to the CEFR. The test taker is indeed seen as the starting point of test development and validation (See Mislevy et al., 2003; 2004; Weir, 2005a; Kane, 2008; O'Sullivan & Weir, 2011). The significance of test taker characteristics in testing has also been underlined by various experts in the field (Bachman, 1994, 2004; Bachman & Palmer, 1996; Khalifa & Weir, 2009; Kunnan, 1995; O'Sullivan, 2000; Purpura, 1999; Shaw & Weir, 2007). However, as supported by the data accumulated in all phases of the research (Chapters, 4, 5 and 6), the CEFR linking process seems to have little impact on the test taker considerations of the examination. In Phase 1, the data collected through field notes (for familiarisation and standardisation) and interviews (for specification and empirical validation) in particular

revealed that test taker characteristics do not play a significant role in the CEFR linking process. In Phase 2, the in-depth evaluation of the process, the same conclusion was drawn regarding test takers. As for Phase 3, the review of the Manual demonstrated that the specification stage of the linking process required users to specify 'who' the test takers of an examination are, such as adults and young learners, but not how their characteristics are dealt with through the test development process. It should be reiterated here that this research imposed a validation theory onto the process stipulated by the Manual. While acknowledging that it is not among the purposes of the Manual to guide test design or validation and thus consider test takers, it is a cause for concern that what lies at the heart of testing, viz. the test taker, was not more in evidence in the CEFR linking process. The significance of test taker considerations was emphasised by O'Sullivan (2000) over a decade ago when rationalizing his approach to defining the test taker.

1b. To what extent does the linking process guide those undertaking a linking study to focus on the context validity of an examination?

The linking process had the greatest impact on the context validity of the examination. Phase 1 revealed that task design and demands were frequently discussed at all stages of the linking process. They facilitated a better understanding of what the COPE examination measured. The field notes particularly revealed that the discussions on the parameters of context validity of COPE tasks led to minor revisions, which were later incorporated into the test specifications of the examination. In Phase 2, the questionnaire demonstrated that context validity parameters were again frequently considered throughout the linking process. Finally in Phase 3, it was indicated that especially through specification forms and at the standardisation stage, the users were

forced to clarify the context parameters of the examination, task demands in particular. The examination was not developed based on the CEFR; however, the linking process definitely facilitated a better understanding of what COPE measured, in other words, the performance conditions (cognitive load) and operations (cognitive processing) of the examination.

1c. To what extent does the linking process focus the attention of those carrying out a linking study on the cognitive aspect of validity of an examination?

Phase 1 of the research pointed in the main to one of the parameters of the cognitive aspect of validity, language knowledge, as being most relevant to the CEFR linking process whereas other parameters became more apparent in Phase 2. In the linking process the project members were asked to extrapolate what executive resources are required to complete the test tasks. For instance, in Phase 1 of this research, 'language knowledge' as an executive resource emerged as the most striking parameter. In Phase 2, parameters that are common to most language models of reading such as word recognition, monitoring and purpose for reading and generating ideas, organizing ideas and strategies for writing seemed to receive attention. On the other hand, Phase 3, the review of the Manual, revealed that qualitative analysis methods such as reflection and analysis of samples to investigate the cognitive requirements of an examination were emphasized in the Manual. The researcher found that the analysis of samples by the participants was useful in exploring the cognitive requirements of a test as was the case at the standardisation stage reported in section 4.4.4 and 5.4.3. However, Wu (2011) indicates that expert judgment on its own is not sufficient to investigate how test takers process test tasks.

1d. To what extent does the linking process emphasise the importance of the scoring validity of an examination?

Scoring validity has two dimensions in this research; the scoring validity of the COPE examination as perceived by Weir (2005a) and the scoring validity of the standard setting, that is accumulating evidence on how well the judges performed and how reliably the standard setting was undertaken. The scoring validity of the COPE examination was focused on in detail under empirical validation in section 4.5.4.2, which revealed that with a reliability of .66 for the reading paper and an inter-rater reliability of .69 for the writing paper, the exam was moderately reliable. Analysis of the COPE writing criteria in the same section in Chapter 4 showed that the criteria worked well. As regards scoring validity, the Manual gives guidance on how data, such as the ones gathered for COPE, can be collected. For instance, it recommends users to carry out CTT or IRT analysis and use the reflection method to investigate rater reliability in Chapter 6 (Council of Europe, 2003) and provides guidance on these analysis methods in its Reference Supplement (Council of Europe, 2004) as discussed in Phase 3.

The scoring validity of the standard setting is commonly referred to as internal validity, which is among the types of evidence accumulated to validate a standard setting event (Dawber, Lewis & Rogers, 2002; Kane 1994, 2008), and requires quantitative empirical information about the consistency of the participants' judgments during a standard setting event (Pant, et. al., 2009). In this research, as part of scoring validity, the use of the criteria, CEFR levels and scales in this case, proved to be the most focused on and the most used parameter at all stages except for the empirical validation stage. The criteria in this case are the CEFR levels and scales. Evidence as to the understanding

and use of the criteria was collected at the familiarisation and standardisation stages of the linking process in Phase 1 of the research. Phase 2 revealed that test reliability, item analysis results, type of marking and marking criteria were considered especially in the standardisation stage of the linking process. Accumulating scoring validity evidence regarding the standard setting of COPE proved to be successful in terms of internal, external and consequential aspects. Internal validity evidence of the standard setting showed high agreements among participants in the standardisation stage; external validity in the empirical validation revealed high agreement between teacher judgments and the COPE categorization of students; and consequential evidence, again in the empirical validation stage, demonstrated that the cut scores established were similar to the old boundaries employed for COPE and therefore, did not have any implications on the pass rates.

1e. To what extent does the linking process have an impact on the consequential validity of an examination?

As borne out in all phases of the research, the consequential aspect of validity has only a limited role in the CEFR linking process. To be more specific, at the standardisation stage, the project members discussed the impact of the test and the standard setting results on the institution (Phase 1 and Phase 2). However, this did not result in any collection of data that might have contributed to a consequential validity argument as the impact was to be investigated after all sections in the COPE examination had been linked. The only evidence collected was a comparison of the old boundaries and the CEFR cut scores established during the process, which showed that the boundaries were almost the same, not having an impact on the pass rates.

1f. To what extent does the linking process have an impact on the criterion-related validity of an examination?

The Manual suggests that users carry out 'external validation', which involves using teacher judgments or performance on another test to verify the cut score recommended at the end of the standardisation stage (Council of Europe, 2003). In fact, at the empirical validation stage, comparisons between the COPE examination and other tests such as FCE or CAE had to be made as the exemplars provided by the Council of Europe consisted of sample tasks from these tests (Phase 2). However, the suitability of the exams used for external validation or criterion-related validity needs to be questioned prior to undertaking such a study because for the comparisons between exams to be meaningful, the two tests must be similar in terms of their underlying construct and evidence regarding the validity of the linkage claims is essential as discussed in section 4.5.4.2. The Council of Europe provides reading samples from the Finnish Matriculation Examination, Cambridge ESOL FCE and CAE. However, even though general information on the purpose of these exams is provided online, in-depth information regarding their constructs or design features is not available to users. In the COPE linking study, the target level was B2 but using FCE samples for criterion-related validity was a concern as FCE measures general English and the test tasks are designed accordingly whereas COPE aims to measure academic English (Phase 1). The purpose of an exam has implications on its task types, text genre, linguistic complexity, e.g. pragmatic knowledge, and the type of sub-skills/strategies tested.

Table 7.1 below attempts to summarize the findings of RQ1 by highlighting the areas covered in the CEFR linking process. In the table, the acronyms refer to the stages of the linking process (F: familiarisation; Sp: specification; St: standardisation; EV:

empirical validation). The shaded areas show aspects of validity that were covered at a certain stage in the linking process whereas the areas not shaded show aspects of validity that are not considered in the process. Overall, the table suggests that the CEFR linking process is lacking in its capability to capture a complete validation argument. This issue is further discussed in section 7.4.1.

Table 7.1 Parameters of validity as tackled in the CEFR linking process

	STAGES OF THE CEFR LINKING PROCESS			
	F	Sp	St	EV
TEST TAKER				
Physical/physiological				
Psychological				
Experiential				
CONTEXT VALIDITY				
Task setting				
Task demands				
Administration				
COGNITIVE VALIDITY				
Executive resources (cognitive load)				
Executive processes (Cognitive processing)				
SCORING VALIDITY				
Marker Reliability				
CONSEQUENTIAL VALIDITY				
Differential				
Washback				
Effect on society				
CRITERION-RELATED VALIDITY				
Different versions				
Different occasions				
Other tests				
Future performance				

7.2.2.2 Research Question 2

Is the CEFR linking process equally applicable to tests of reading and writing?

2a. ***What variations, if any, are there in the Manual's methodology to the validation of productive and receptive language tests in terms of attention to test taker considerations?***

As explained under 7.2.2.1, the importance of test taker characteristics has been discussed by specialists in the field of testing and the role of individual characteristics for reading and writing in test development and production has also been studied (Khalifa & Weir, 2009; Kunnan, 1994, 1995; Shaw & Weir, 2007). It was also indicated in the same section that the Manual methodology does not take the test taker into consideration. We could therefore state that the answer to the above research question is clearly that the Manual methodology does not contribute to the validation of either productive or receptive language tests in term of test taker characteristics.

2b. ***What variations, if any, are there in the Manual's methodology to the validation of productive and receptive language tests in terms of attention to context validity?***

The contributions of the Manual methodology to the validation of productive and receptive papers were similar in terms of context validity, in that all stages of the process require users to carefully consider the task design and task demands in particular. Both of these aspects of a test are questioned in the same way in the specification forms. However, reading and writing skills entail different cognitive processes and therefore, need to be examined differently. For instance, due to the nature of these skills, productive and receptive, the strategies required in both differ in some

aspects. The CEFR clearly states that although both productive and receptive skills require planning, execution, evaluation and repair strategies, it differentiates what these strategies mean for productive and receptive skills (Council of Europe, 2001). Whereas evaluation for writing involves monitoring success, evaluation for reading involves hypothesis testing and matching cues to schemata. Similarly, execution for writing involves compensating, building on previous knowledge and experimenting while it requires identifying cues and inferring from them for reading. Therefore, although the Manual refers users to the CEFR itself for these skills at the specification stage, it could have adapted a different approach to examining or defining these skills in a test. An example of a different approach might be providing the relevant parameters in the specification forms designed for each productive and receptive skill. For instance, the written production form could include a list of possible tasks and activities for users to choose from. This would also address the problem experienced in the BUSEL study where participants, during the specification stage, had difficulty in differentiating 'tasks' from 'activities', terms used in the forms and the CEFR. The project members had to carefully analyse the examples provided in the CEFR to be able to identify the difference between these terms in order to complete the forms.

*2c. **What variations, if any, are there in the Manual's methodology to the** validation of productive and receptive language tests in terms of attention to cognitive validity?*

As regards test taker characteristics and context validity parameters, the impact of the Manual methodology on writing and reading tests were similar. In terms of cognitive validity, on the other hand, the impact showed some variations. Phase 1 of the research showed that language knowledge, a parameter of the cognitive aspect of validity, was

taken into consideration for both. In Phase 2, as reading and writing have some different (e.g. monitoring, word recognition for reading and organizing and revising for writing) and some common parameters (e.g. textual and background knowledge) due to their nature, they were analysed separately. Data was in evidence that amongst the common parameters, purpose, sub-skills and strategies played a more important role in reading than they did for reading, as was seen in section 5.4.3.2. Finally, data gathered in Phase 3 revealed that the cognitive aspect of validity is dealt with through the specification forms, which force users to consider the cognitive processes required in a task. The users are also encouraged to collect data in this regard, that is the thought processes of test takers, using qualitative methods in the empirical validation stage. However, as mentioned above in 2b, reading and writing involve different cognitive processing and therefore, might require different kinds of analyses.

2d. ***What variations, if any, are there in the Manual's methodology to the*** *validation of productive and receptive language tests in terms of attention to scoring validity?*

In terms of scoring validity, the contribution of the Manual to the productive and receptive tests was slightly different. In Phase 3, it was found that the Manual proposes that users provide evidence regarding the quality of an exam, referred to as 'internal validation' (Council of Europe, 2003: 102). It also indicates that 'a theory as a general framework' is needed to summarize and evaluate data (ibid); however, what is meant by this is employing CTT or IRT analysis, it does not refer to a validation theory. However, the suggested methods are only applicable to the reading paper. In terms of the scoring validity of the writing paper, the Manual offers limited guidance (which refers to the use of reflection or protocol methods) to investigate how raters perform

marking. Recently an addendum to the Reference Supplement of the Manual has been made on Multi-Facet Rasch analysis (Eckes, 2009), which should address this problem.

As mentioned in answering research question 1d, another dimension to scoring validity exists in a linking study; the validity of the standard setting. In this respect, the Manual contributes similarly to both skills through analysis of judge agreement and consistency and also on how to set cut scores, in that it recommends different types of analysis that could be undertaken and provides guidance on how to carry out such analysis. However, the emphasis is again on the methods used for receptive skills, as borne out in Phase 3. Productive and receptive skills, however, require different methods of standard setting and methods for productive skills are not explored in any great detail. Phases 1 and 2, on the other hand, pointed to a different parameter, assessment criteria, in scoring validity. The CEFR scales are seen as the criteria in the linking process; therefore, a lot of analysis of the scales took place in this project mainly due to the need to have a good understanding of the CEFR scales and levels; and to be able to apply them to the examination under study.

2e. ***What variations, if any, are there in the Manual's methodology to the validation of productive and receptive language tests in terms of attention to consequential validity?***

The contribution of the Manual to receptive and productive tests in terms of consequential validity is similar, though the evidence from the research suggests that in both cases this contribution is minimal, as supported by data in all Phases of the research. However, it would appear from the evidence presented here that the linking process itself might have some impact on the participants, the test being linked and on

the classroom. With regard to the participants, as a result of the discussions undertaken, they appeared to become more aware of the meaning of a particular level and how BUSEL students might be raised to that level (taking into consideration both teaching and assessment). In terms of the test, the process might well highlight the need for modification in general or at the task, text, or item level, so that it better reflects the intended level. Similarly, in a school context, there may be implications for the classroom. For example, during the standardisation stage in the BUSEL linking study, the participants identified certain weaknesses in students' responses to the writing prompts that hindered them from reaching the B2 level. The effect of this was a decision to slightly modify the methodology of teaching writing in the school as mentioned in section 6.4.1.5.

2f. ***What variations, if any, are there in the Manual's methodology to the validation of productive and receptive language tests in terms of attention to criterion-related validity?***

In terms of criterion-related validity, a number of suggestions are made in the Manual. For example, an IRT item bank can be set up to ensure parallel forms of an examination. The test can be calibrated to another test which has already been linked to the CEFR, or teacher judgments can be used to confirm the recommended cut scores set during the standardisation stage. Data from all stages of the research showed that the linking process helped collect criterion-related evidence for COPE. However, the procedures suggested by the Manual are of value mainly for receptive skills, though teacher judgments can also be used for productive skills. More guidance is needed in the Manual on how criterion-related validity evidence for productive tests might be

established. One such approach, for example, is to send scripts to external experts, a strategy successfully used in Phase 3 of this study.

A possible concern here involves the findings related to research questions 1 and 2. One might argue that the findings of the research are mostly restricted to what aspects or parameters of validity are considered and discussed throughout the CEFR linking process and do not necessarily suggest that the process in fact contributes to the validity argument of an examination. However, as demonstrated so far, the Manual approach deals with certain parameters of validity, which can suggest that the Manual encompasses an implied view to validity, albeit restricted and outdated (O'Sullivan, 2009a; O'Sullivan & Weir, 2011). If systematically documented, as suggested in section 6.4.1.3, evidence can be accumulated supporting some of the aspects of validity. A suggested approach for this is presented in section 7.4.1.

7.2.2.3 Research Question 3

To what extent does the CEFR linking process help test providers to establish an appropriate level for a test?

3a. How does the linking process contribute to the understanding of the standards set through the examination?

In Phase 1, the answer to this research question came from three stages of the linking process; specification, standardisation and empirical validation. At the specification stage, the interview data revealed that the process of completing the specification forms facilitated understanding what the COPE examination measured mainly due to careful consideration of the different competences involved in language assessment, such as

pragmatic competence or socio-linguistic competence. At the standardisation stage, all respondents to the questionnaire indicated that undertaking the process of analysing and discussing the tasks and items contributed to their understanding of the level of both the reading and writing papers of the COPE examination. This finding was also corroborated through the analysis of field notes kept at this stage. At the empirical validation stage, the analyses carried out revealed that the COPE examination was in line with the Cambridge B2 level, although concerns arose regarding FCE items as discussed in 4.5.4.2, and that the level set through the examination was a challenge for students, in that, many students were not at B2 level.

Data gathered in Phase 2 of the research also supported the finding that the specification stage and, more significantly the standardisation stage, facilitated a better understanding of what the COPE examination measured, its level and problematic areas of the examination that require revision.

3b. Can the linking process suggest ways in which an examination could be modified to raise the level of the examination to pre-determined standards?

The CEFR linking process helped identify aspects of the COPE examination that could be modified to bring the level of the examination up to pre-determined standards. In particular, Phase 1 pointed to a number of areas in this respect. For instance, at the specification stage it was mentioned that the categories provided in the forms (such as socio-linguistic competence, pragmatic competence or strategic competence) encouraged users to consider different aspects of the assessment of reading and writing. In cases where these aspects turned out to be at a lower level than expected, it again forced users to understand the reasons behind what was a jagged profile in terms of the

level of the different skills and competences measured in the examination. It was mentioned by the interviewees that the categories in the specification forms could help carefully specify the level of an examination in all aspects of language ability. The standardisation stage, as supported in Phase 2, contributes most to bringing the level of an examination up to the desirable standard as it involves close scrutiny of items and tasks. This was recognized by O'Sullivan (2009a, 2009b, 2009c) who introduced an extra stage to the CEFR linking process in his alternative model for linking as discussed previously in sections 2.5 and 5.4.7.1. This stage, though similar to standardisation, did not require judges to assign levels to items or performances but critically evaluate them in order to ensure that the items measure at the intended level.

7.3 Limitations

One of the limitations of this study is that since the publication of his book "Language Testing and Validation" in 2005, Weir has updated his validation frameworks for reading and writing in some respects (See Shaw & Weir, 2007 and Khalifa & Weir, 2009). Since the project started in 2006 and the main body of this study was undertaken by the time the first updated model (Shaw & Weir, 2007) was introduced, this research took the original framework as the basis. The reading model was introduced at a later date (Khalifa & Weir, 2009). However, since the only significant changes in the models are in the area of cognitive processing as part of cognitive validity the findings of this study can still be seen to contribute significantly to our understanding of the benefits and limitations of the linking of an examination's cut scores to a set of recognized standards.

Another limitation regards the fact that the study is context-specific as it was carried out

in the BUSEL Preparatory Program. Further studies in linking examinations to the CEFR should investigate the areas explored in this research. The study was also restricted by the daily operations of the school, which mainly affected the research time framework and the design due to institutional factors. For example, there were long intervals between some of the sessions as daily work commitments made it impossible for project members to meet regularly or as frequently as desired. The project members had concerns over whether the videos recorded of their discussions could be screened by the senior management, which led to ethical concerns regarding the use of videos in the study. As a result of these concerns, the use of video recordings was abandoned. Some information related to the COPE examination was confidential; thus certain parts of the research, such as the interview in Phase 3, only involved people who had access to confidential data. However, despite the limitations associated with the study being carried out at a single institution (and the issues arising that have been summarized here), the breadth of the study has contributed significantly to our understanding of both the process and its impact on the institution. In that respect, while some findings may turn out to be institution-specific, many will not be and the experiences outlined in this study will help in future projects where institutions are attempting to gather evidence to make meaningful claims about their examinations with regard to level.

A final limitation is that three years after the commencement of this study, the final version of the Manual was published by the Council of Europe. This does not pose any major implications on this research as in the final version of the Manual the stages of the linking process and their procedures remained the same. The Manual now offers more systematic guidance on all stages of the linking process. In terms of its approach to validation, it has broadened its approach; however, still with the traditional internal

and external validity concepts. Internal validation involves conducting pre-testing, content validity as well as procedural and internal validity of the standardisation stage (Council of Europe, 2009). As for external validation, the Manual incorporated new suggestions that will help gather criterion-related validity. However, since internal validation of the standardisation stage has been addressed under scoring validity in this thesis and external validity was incorporated into criterion-related validity, the changes in the final version of the Manual in terms of its approach to validation are not seen as negatively impacting on the research reported here. In addition, since the latest version of the Manual does not recognise the need for an underlying model of validation, it is still significantly limited in its long-term value.

7.4 Contributions to the field

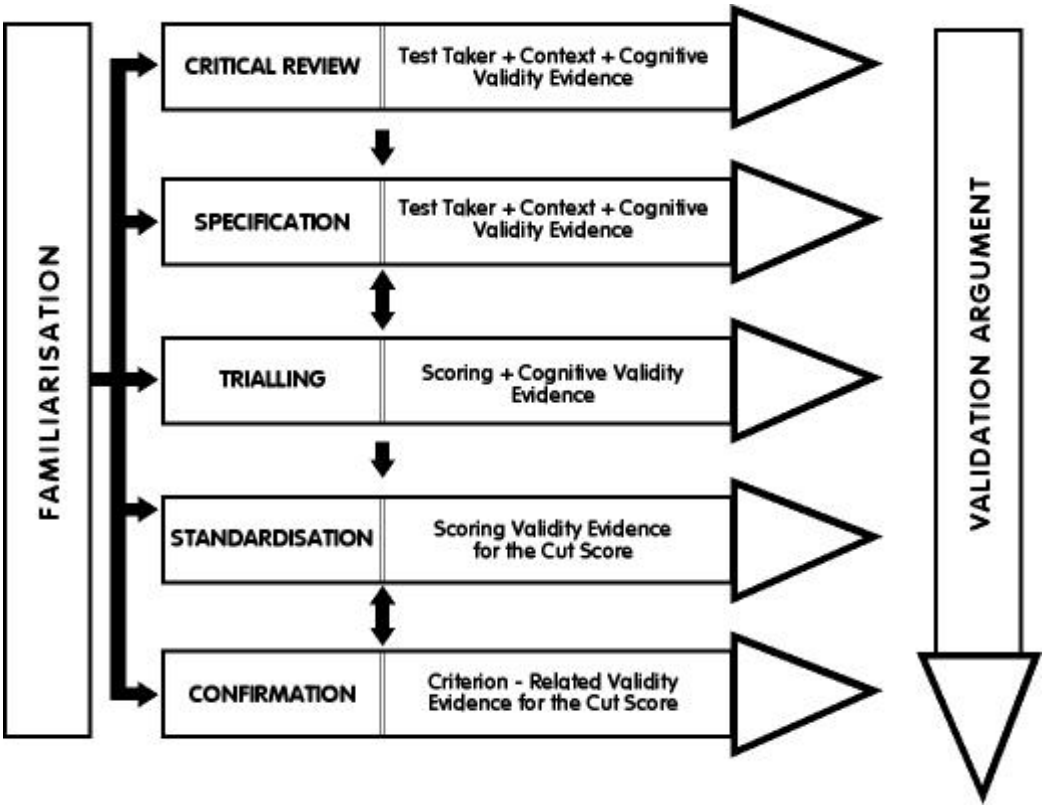
7.4.1 A validation model for linking examinations to external criteria

One of the main findings of this research is that evidence regarding certain aspects of validity is captured through the Manual, though in a rather limited way; however, the Manual does not offer guidance in establishing a validity argument for the examination under study. Different stages of the CEFR linking process provide information about an examination leading into a validation argument. However, connections between the different stages are weak, or in fact do not exist at times. At this point, an important contribution of this thesis to the field of testing, particularly for users of the Manual, regards the validation model in the Manual.

The suggested validation model sets out to accommodate for designing a new examination in relation to the CEFR and for linking examinations that are already in place to the CEFR. It was inspired by the alternative model for linking a test to the

CEFR proposed by O'Sullivan, resulting from his experience with the City and Guilds Communicator Exam linking project (2009a). Figure 7.1 presents an overview of the validation model for linking which takes into account the findings of this study and the model proposed by O'Sullivan.

Figure 7.1 A Validation Model for the Linking Exams to the CEFR



The model follows the Manual's philosophy regarding the familiarisation stage, in that it is an ongoing process and has to be repeated prior to every stage of the linking process. The model starts with a critical review as suggested by O'Sullivan (2009a) "to ensure that the test is working well and has the attributes that will make any linking meaningful" (p. 83), but it will also serve as a check mechanism to analyse an examination in relation to the CEFR and its construction principles prior to the linking process. It is strongly suggested that external participants are invited, particularly at this stage to 'avoid insider's bias' (Papageorgiou, 2007b: 297). This is in fact recommended at all stages of the process. Structured accumulation of data at this stage will provide evidence regarding test taker characteristics, context and cognitive aspects of validity evidence for an examination. Structured accumulation of data may involve keeping field notes, using questionnaires or asking judges to write down their justifications for and/or

comments at item, task and text levels. In addition, data regarding the criteria designed to be used in the writing section, for instance, should also be collected at this stage even though it might be limited to expert opinions only.

The specification stage should be conducted by a group including people who were not involved in the design of an examination, similar to the critical review, to guard against insider's bias. Specific examples taken from the examination in question to support the claims made throughout the specification forms could generate evidence, which would contribute to an argument in support of the test taker, context and cognitive aspects of validity. It should be highlighted once again that the specification forms need to be further developed to lead users in the right direction with regard to test validation. For instance, the criteria designed to be used for rating the writing samples should be examined again at this stage with respect to how they reflect the intended outcomes. To make this process easier, elements of the CEFR need to be incorporated into actual test specifications.

Once an examination is described through CEFR specifications or test specifications, it needs to be trialled before proceeding to the other stages of the CEFR linking process. At the trialling stage, data regarding context, cognitive and scoring aspects of validity can be collected and any design problems that were not noticed at the critical review stage can also be identified Differential item functioning or bias analysis should also be carried out at this stage. However, if an existing test that has been shown to be well-designed and developed is to be linked to the CEFR, additional trialling may not be necessary.

At the standardisation stage, procedural, internal and to a certain degree, external evidence regarding the validity of the standard setting can be gathered. Procedural evidence involves documenting how well the stage was carried out and internal evidence entails data supporting the reliability of the judgment process while setting the cut score and inter-rater reliability of the judges. External evidence through the employment of different standard setting methods can be collected at this stage.

The last stage, confirmation, requires collecting data to confirm the cut score established as a result of the standardisation stage. This can be done through teacher judgments, which is in fact an examinee-based standard setting method, and comparing the exam with an external test that claims to be measuring the same abilities. The confirmation stage in fact adds to the external evidence for the validity of the standard setting.

The evidence gathered from all stages of the suggested model will contribute towards putting forward a full validation argument. It is acknowledged; however, that evidence accumulated at the critical review, specification and trialing stages is not sufficient to address context and cognitive aspects of validity. Further work needs to be carried out in these areas for an enhanced validation argument. (See Weir et al. (2006; 2008) and Moore et al. (2008) for studies exploring context and cognitive aspects of a reading test).

7.4.2 **Adaptation of Weir's validation framework**

The research also brought a new perspective to validation models in terms of the role of standard setting. Standard setting involves classifying learners into a number of proficiency levels such as "advanced", "proficient" and "basic" (Kane, 2001: 53) by

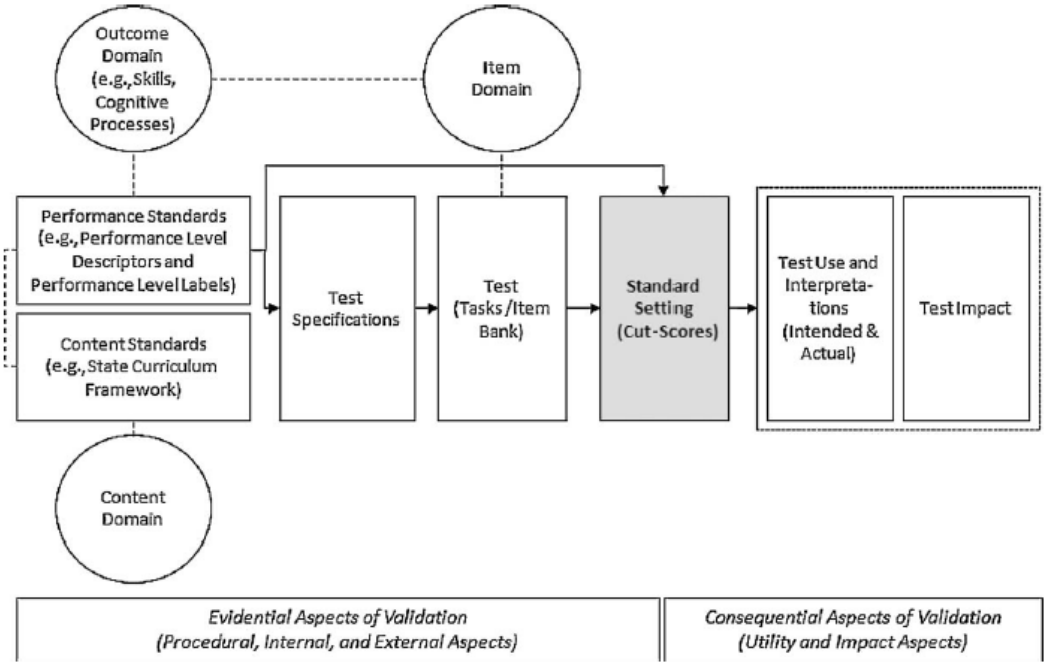
determining a critical score on the test that draws the line between sufficient and insufficient performance on a given test and sufficient performance is defined in relation to a given purpose (Zieky et al, 2008). The role of standard setting in validation can be analysed in two ways: standard setting in CEFR linking and standard setting as part of test development.

In the Manual, standard setting is presented as a stage in the linking process. Guidance and procedures regarding standard setting are only provided in the standardisation stage of the linking process, Chapter 3 of the Manual. However, standard setting is, in fact, an integral part of the whole validation and development process. At all stages of the CEFR linking process, the focus was on the CEFR, the external criterion used to link the COPE examination to. In other words, the CEFR was the standard COPE was set against. At the familiarisation stage, the project members strived to learn the CEFR, its levels and particularly the requirements of the B2 level so that they could use this knowledge in the subsequent stages of the linking. At the specification stage, the exam content was analysed in relation to the CEFR to ensure that the requirements set for the B2 level were captured in the COPE examination. At the standardisation stage, reading items and written scripts were analysed with respect to the CEFR levels so that the B2 cut score could be set. Finally at the empirical validation stage, links to items which were claimed to have been calibrated at the B2 and C1 levels in addition to teacher judgments were established to confirm the cut scores. When these are considered, it can be claimed that linking, with all its stages, is in fact standard setting. In addition, as summarized in 7.2.2, each stage of the linking process has close links with certain aspects of the validity and validation of an examination, so linking itself is actually an aspect of validation in its broad sense. However, standard setting in its traditional sense

has not been emphasized in validation frameworks, which brings the discussion to the second part of the analysis of standard setting as part of test development.

Standard setting is seen as a critical component of test development because "unless cut scores are appropriately set, the results of the assessment could come into question" (Bejar, 2008: 1) also because standard setting is critical in interpreting learner performance (Cizek & Bunch, 2007; Pant et al., 2009; Zieky & Perie, 2006). Pant et al. state that "standard setting studies are a critical gateway between *evidentiary aspects* of validation and *consequential aspects* of validation" (2009: 97) and present this relationship as shown in their diagram given in Figure 7.2. It should be noted that this model does not intend to present a validation model but aims to demonstrate the role of standard setting for evidentiary and consequential aspects of validation. The role of standard setting in validation models should in fact be examined.

Figure 7.2 The role of standard setting for evidentiary and consequential aspects of validation by Pant et al. 2009: 97



The most comprehensive validation models, such as those of Weir or Mislevy, aspire to analyse an examination from almost all angles as presented in detail in section 2.4.2. At the design stage, validation models encourage testers to focus on the language theory behind a test, thought processes of test takers, task types and administration. After its administration, the focus switches to marking and scoring, and evidence is collected to support the reliability of the test and its marking. As a final stage, validity evidence regarding the areas considered at the design stage such as the intended thought processes of test takers is collected while *a posteriori* evidence is accumulated, as Weir suggests (2005a). Such validation studies aim to provide evidence regarding the claims made at the design stage with respect to what a certain test sets out to measure and at what level. Whereas validation frameworks examine the stability and accuracy of the standards, they do not seem to be exploring issues related to the initial standard setting of a test. For example, they do not appear to be investigating why the standard for a given test was set at a certain level, how it was set and what measures were taken to provide evidence regarding the plausibility of that standard. In fact, leaving the checking of the level or the standard to the end or until after a test is administered might be costly for institutions because of possible implications resulting from a faulty cut score, i.e. if the level agreed on at the pre-development stage is not the level actually required. For instance, in school contexts, students may be misplaced by a test, not because the test fails to measure the desired abilities but because the cut score was not set at the appropriate level. This argument is valid for examinations that employ internally set standards and those that use external criteria to set standards such as linking examinations to the CEFR. In fact, in the case of using external criteria such as the CEFR, alignment studies involve aligning the internally set standards of a test with respect to an external standard.

As part of this contribution, questioning the role of standard setting in validation, the need to theorise and update existing models arose. In order to address the issue of standard setting, the diagram in Figure 7.3 aims to present a validation model, an adapted version of Weir's framework. In this new model, standard setting is seen as the parameter that integrates *a priori* aspects of validity with *a posteriori* aspects. The interaction among test taker characteristics, context and cognitive elements that form construct validity (O'Sullivan & Weir, 2011) are considered to be *a priori* and scoring, consequential and criterion-related aspects of validity are seen as *a posteriori* by Weir (2005a). Although the conceptualisation of the standard comes from the design stage, standard setting is the last stage of test construction as it is the confirmation that the cut score reflects the agreed standard. As such, in the model, standard setting is seen as the core of post-test analysis. In developing tests with a view to validation right from the design stage, a test is constructed with a particular candidature and their needs in mind (test taker characteristics), which form the purpose of the test. The tasks are chosen to ensure that the required cognitive processes are measured in the right way (context and cognitive aspects). Once the exam is ready, the expectations need to be revisited to set a suitable standard, in other words to establish the cut scores that will determine pass/fail decisions. This is where standard setting comes into play. The standard setting in itself should be validated by collecting procedural, internal, external, and consequential evidence (Dawber, Lewis & Rogers, 2002; Kane 1994, 2008). These are briefly explained in the Table 7.2, adapted from Pitoniak (2003).

Table 7.2 Standard setting validation parameters

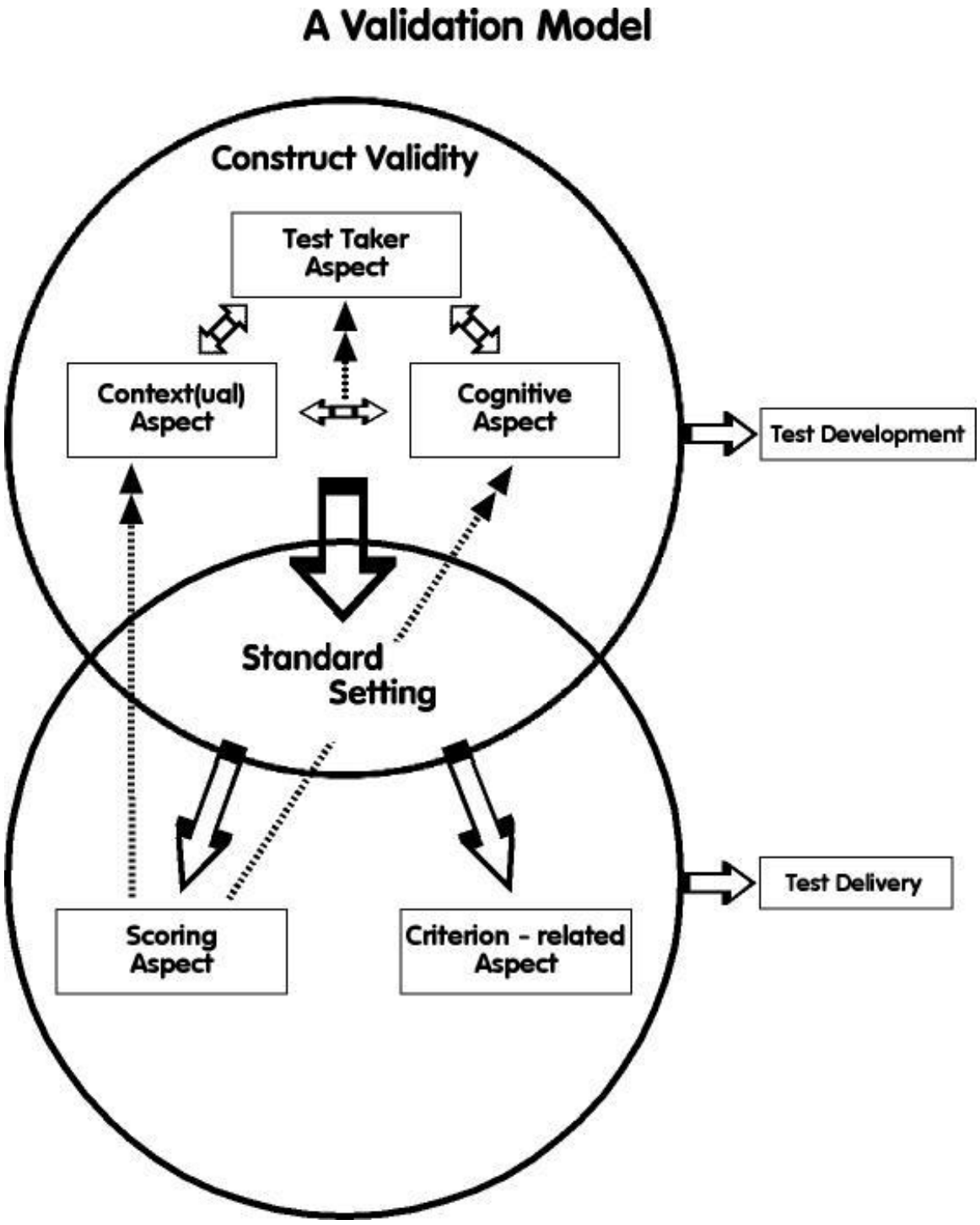
Parameters	Details
Procedural	Explicitness Practicability Implementation Feedback Documentation
Internal	Consistency within method Intra-participant consistency Decision consistency Other measures (consistency of cut scores across item types, content areas, and cognitive processes)
External	Comparisons to other standard setting methods Comparisons to other sources of information
Consequential	Reasonableness of cut scores Adequacy of reporting and reception

After the test is administered, the standard, or the level set is inspected in two ways. Firstly, investigating the scoring aspect of validity sheds light on whether the standard is perceived and applied accurately by others in the case of productive skills and whether the test version is constructed as it is supposed to have been in the case of receptive skills. This parameter of the scoring aspect of validity can be referred to as 'level internalization'. Evidence regarding criteria and marker reliability can be collected for productive skills and items analysis can be carried out for receptive skills. Secondly, evidence of the criterion-related aspect of validity not only helps confirm that the intended standard is actually realized, for instance through comparison with another examination, but it also contributes to the process of maintaining the standard initially set.

Two issues might raise questions regarding this new model of validation. Unlike Weir's framework, each component in the model is referred to as an 'aspect' of validity. It has

been long accepted that validity is a unitary concept and that there can only be a validation argument consisting of different types of evidence resulting from different aspects of a test. Therefore, the word 'aspect' might be more suitable to use while talking about the components that make up a validation argument. In addition, in the validation model presented here consequential validity is omitted. The place of consequences in a validation argument has been controversial for decades (Cizek, 2011). Whereas the importance of consequences has always been acknowledged, its inclusion in validation frameworks is seen as an error by various specialists in the field (Cizek, 2011; Shepard, 1997; SIOP, 2003). This view, a position long taken by O'Sullivan (2011), has also led to a recent reconceptualisation of Weir's validation framework. (O'Sullivan, 2011; O'Sullivan & Weir, 2011). O'Sullivan and Weir (2011: 8) advocate that "an ethical approach to test development is a reflection of an understanding of the consequences of decisions made during the process of development". They continue to suggest that without a firm understanding of test construct with respect to test takers, context, cognitive demands and the scoring system, it is not possible to adopt an ethical approach to test development. In other words, although the impact of a test needs to be explored, it should be kept separate from a validation argument because, as Cizek argues, "evidence gathering to support intended interpretation(s) is distinct from evidence gathering in support of a specific use" (2011: 17). A test is developed for a specific purpose and if used for an aim other than the intended one, then the consequences should be investigated in order to determine whether that test is also suitable for that secondary purpose. However, in this case, it is not only the responsibility of the test developer to collect evidence but of the institution that decides to use the test for the alternative purpose as the institution has to ensure that the test provides a valid measurement for the alternative purpose.

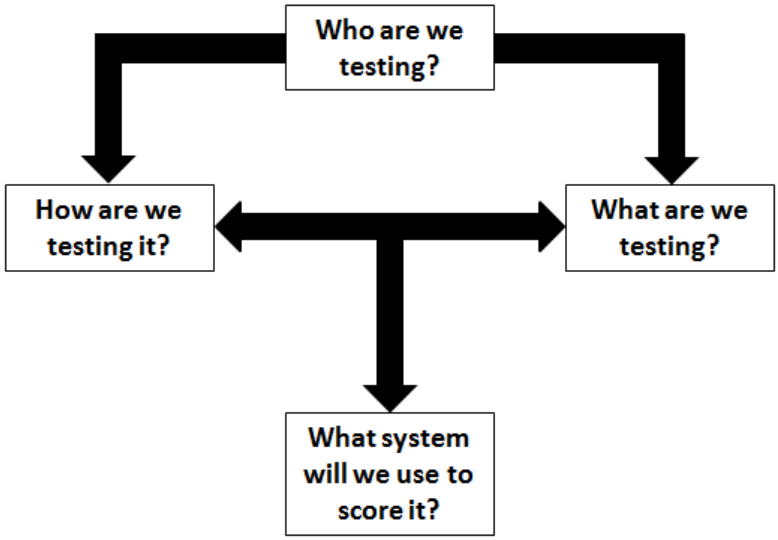
Figure 7.3 **An adaptation of Weir's validation framework**



7.4.3 Definition of construct validity

Weir regards context and cognitive aspects of validity together with the scoring criteria as lying "at the heart of construct validity" (2005a: 85) and this idea was further developed by O'Sullivan and Weir (2011). According to the new definition, construct validity encompasses the test taker, context, cognitive and scoring aspects of validity. They present the key interaction between these aspects of validity and the key questions as seen in Figure 7.4.

Figure 7.4 **The 'core' of construct validity by O'Sullivan** and Weir (2011)



In this research, evidence related to the interaction between these aspects of validity has come to light. This evidence was found in the field notes kept during familiarisation and standardisation, the interview carried out at the specification stage and the Phase 2 questionnaire and pointed to areas that are common in particular to the context, cognitive and scoring aspects of validity. For example, criteria for marking writing, a part of scoring validity, ideally reflect the task demands of a test and the cognitive resources measured through the task. The participants in this project had to talk about the CEFR scales, which are the criteria used to categorize local writing samples and

discussed the requirements of a task and language resources needed while assigning CEFR levels to sample performances. This natural tendency to refer to context, cognitive, and scoring aspects of validity revealed the close link among the parameters that are the core of construct validity.

At this point, the role of the specification stage of the CEFR linking process gains importance. The specification stage needs to be further developed through improving the procedures for the completion of the forms and the forms themselves. The specification forms should be completed by a group of people including the designers of the test under scrutiny as well as its users (such as teachers and external participants, as mentioned above).

The specification forms should be redesigned so as to clearly reflect aspects of construct validity. This could enable those who carry out CEFR linking projects to scrutinize their examinations, guide their revision if the need arises so that they better reflect the intended levels and attributes; and most importantly, accumulate evidence towards construct validity. This type of evidence is not only collected at the specification stage, it can also be gathered through keeping records of the discussions held throughout the standardisation stage because, as was presented in various sections of this thesis (Chapters 4, 5, and 6) in justifying the levels they assigned for individual items in a test, the participants of a linking project reflect on most of the parameters of construct validity at the standardisation stage. As mentioned in section 7.4.1, further evidence from test takers themselves should also be accumulated to have an enhanced validation argument. See Weir et al. (2006; 2008) and Moore et al. (2008) for studies exploring context and cognitive aspects of a reading test.

7.4.4 Test quality

In a study which aimed to investigate judgment-making in the CEFR linking process, Papageorgiou (2007a, 2007b) stated that "research needs to provide a better understanding of the kind of impact the CEFR linkage might have on test quality". This research has shed light on the issue by providing evidence towards the contributions of such a study on the validity argument of a test, which is the core of test quality. As demonstrated in this research, linking an examination to an external criterion such as the CEFR facilitates the validation process if carried out in a principled way, as suggested in section 7.4.2 above. The specification stage of the linking process calls for a rigorous scrutiny of the exam under examination in all aspects, from construct to administration. Should the specification forms be improved, this stage has the potential to have an invaluable impact on test quality. The standardisation stage involves standard setting and its validation leading to a feasible cut score. The empirical validation stage enhances the validation argument from a different angle; criterion-related validity.

7.4.5 Problems encountered throughout the CEFR linking process

One of the most valuable contributions of this study is that it reflects on the experiences surrounding a linking study.

Table 7.3a Problems encountered throughout the CEFR linking process (Project setup)

Stage of CEFR linking	Problems encountered	How problems were dealt with throughout the project
Setting up the project	Difficulty in drawing up a realistic project framework	A close analysis of the Manual was carried out but problems could not be completely anticipated so articles outlining the problems others have encountered were utilized
	Identifying resources required before the project started	Dealt with resource issues as the project proceeded

The study provides a record of all the problems encountered throughout the study, some of which result from a lack of guidance on the part of the Manual and some others are due to the CEFR being underspecified. A list of these problems is given in Tables 7.3a to 7.3e. The lack of guidance in the Manual caused problems for the running of the linking process whereas the problems originating from the CEFR itself made it difficult to carry out every stage of the process. The table also summarises how these problems were dealt with in the project.

Table 7.3b Problems encountered throughout the CEFR linking process (Familiarisation)

Stage of CEFR linking	Problems encountered	How problems were dealt with throughout the project
Familiarisation	Working with an inexperienced group who were not all familiar with the CEFR	An extended familiarisation stage was conducted. Carrying out the Manual familiarisation activities or working with the CEFR itself was not sufficient to familiarise participants sufficiently with the CEFR. Articles exploring both positive and negative sides of the CEFR, others' experiences in using the CEFR etc. were used.
	Complexity of understanding the CEFR itself	An attempt to further define the descriptors as a group was made to relate them to the context
	Difficulty in understanding what the linking process involves	Relied on others' experiences and the experiences of some the project members with similar studies
	Difficulty in understanding how the CEFR descriptors can be used in different contexts e.g. academic	Further defined the descriptors to relate to the context through group discussions of what the descriptors meant for the academic context
	Restricted usefulness of the Manual familiarization activities	Made use of locally designed tasks i.e. quizzes
	CEFR global and overall scales focus mostly on language knowledge, not enough emphasis on cognitive processing, which makes it difficult to conceptualise the CEFR level expectations	Further defined the descriptors to understand the cognitive processing involved through group discussions
	Lack of guidance on how to check/monitor familiarity – judging when to move on	Conducted a number of statistical analyses e.g. MFR

Table 7.3c Problems encountered throughout the CEFR linking process (Specification)

Stage of CEFR linking	Problems encountered	How problems were dealt with throughout the project
Specification	CEFR is underspecified leading to difficulties in analyzing exam tasks and filling in the forms	Group discussions to clarify the task requirements
	Manual suggesting that the forms be filled in by exam developers – outsiders perception is required to avoid institutional bias	The group included people from different parties in the school e.g. teachers, textbook writers, etc. to avoid bias
	Lack of guidance and lack of an example make the completion of forms difficult	The forms had to be revised until they reached a satisfactory level of detail and could be interpreted meaningfully
	Lack of guidance on how judge the quality of the form completion process	High levels of group agreement on these issues were sought for before we considered moving on
	Language competence forms most difficult to fill in as CEFR is underspecified	Had to be revised several times and justifications had to be made for the jagged profile
	Lack of guidance in interpreting the outcome of the specification stage – the graphical representation of exam levels in relation to the CEFR	Group interpretations and justifications were made regarding the outcome
	Lack of focus on test taker considerations and context resulting in a gap between the exam and the CEFR	Target language situation was considered and discussed throughout the project

Table 7.3d Problems encountered throughout the CEFR linking process (Standardisation)

Stage of CEFR linking	Problems encountered	How problems were dealt with throughout the project
Standardisation	Limited range of sample calibrated items available – not always relevant to context	DIALANG samples were used for training and local samples were sent out to external experts for a second opinion
	Lack of precise / detailed information regarding the calibrated items – e.g. FCE items all at the same level, not distinction made regarding the level of items within a CEFR level	Used a variety of samples from the same exams and administered in different tests. The results were compared but were inconclusive
	Lack of guidance on how to apply CEFR scales to test conditions	Formulated own ground rules
	CEFR descriptors underspecified causing judgment making difficult	Descriptors further defined for the context

Table 7.3e Problems encountered throughout the CEFR linking process (Empirical Validation)

Stage of CEFR linking	Problems encountered	How problems were dealt with throughout the project
Empirical validation	Lack of a validation theory adapted in the Manual hinders a strong validity argument for an exam	Weir's validation framework was adopted
	Empirical validation seen as the last stage of linking whereas it has to be done throughout	Validity evidence was collected throughout the project

The issues highlighted above can be summarised as:

- There is a serious lack of practical advice in the Manual related to how a linking project might be set up and run.
- The lack of sufficient detail in the CEFR level descriptors means that any

organisation planning to run a project such as that outlined in this thesis first needs to consider holding extensive discussions within the organisation, which lead to further local detailed definition of the interpretation of the CEFR levels. That is, to carry out this type of study, both domain and context specific interpretation of the CEFR is required.

- The lack of clarity in the Manual of the role of the specification forms in the linking process means that it is very difficult to know what information is being sought out (and why) and equally difficult to know how it is to be interpreted.
- The Council of Europe recommended exemplar items are weak in terms of both calibration and focus. With regard to the former, there is insufficient psychometric information to allow for their use as standardised items, while their language focus is limited and unlikely to be useful in a non general-proficiency context.
- The view of validity and validation in the Manual is limited and unhelpful. The importance of theory to the process of linking is seriously lacking.

7.5 Implications

A number of implications have emerged from the research, mainly for examinations, the Manual and models of validation.

7.5.1 For examinations to be linked

One of the research questions explored in this thesis was "To what extent does the CEFR linking process help test providers to establish an appropriate level for a test?" and it had two sub-questions:

a. How does the linking process contribute to the understanding of the standards set through the examination?

b. Does the linking process help identify aspects of the examination that could be modified to bring the level of the examination to pre-determined standards?

While summarizing the findings related to these questions in Section 7.2.2, it was stated that the CEFR linking process indeed contributed to our understanding of the standards set through the COPE and pinpointed aspects of the examination that could be modified to bring its level of the examination to a pre-determined standard. Gaining a better understanding of the standards set through an examination is invaluable to test developers and item writers for two reasons. The first reason is that it would contribute to the validity of an examination by amplifying the construct definition with a focus on task demands and level. Based on the experience reported here, at the specification and standardisation stages, the examination tasks and items can even be fine-tuned based on the resulting solid construct definition as was the case with the COPE examination. The test specifications were revised and the writing task was modified so that the test better reflected the intended standards meticulously. This need was acknowledged in the City and Guilds CEFR linking project thus leading to an extra critical review stage prior to standard setting (O'Sullivan, 2009a, 2009b, 2009c). The second reason pertains to maintaining the standards set through an examination. Certain approaches, such as calibrating the examination under study with an external examination already linked to the CEFR or using teacher judgments to set and/or confirm the levels, are valuable tools to maintain its level. Teacher judgments and calibration are now an integral part of the assessment system in BUSEL and in similar contexts such as language preparatory programs at universities, these tools can be used as check mechanism for test level, even

in the case of institutions with a series of examinations like BUSEL where all five level tests are to be linked to the CEFR and calibrated.

7.5.2 For the Manual

Two main implications have emerged from this research for Manual users. Firstly, the research identified areas for improvement with respect to validation. As O'Sullivan (2009a) argued, the Manual approach to validation is limited and outdated and it does not acknowledge advances from Messick to Weir. The Manual does not aim to offer guidance on test development or validation. However, since it sets out to suggest ways to provide evidence as to the validity of the test itself and the linkage claims, users should adopt a validation theory from the beginning of the study in order to provide a complete validation argument. This would be more valuable for an examination than accumulating desultory validity evidence.

Secondly, throughout the COPE linking project it was found that linking to the CEFR was an iterative process, thus supporting the arguments made by O'Sullivan (2009a). The reading standardisation had to be done three times for instance, until a viable cut score could be established. O'Sullivan (2009a, 2009b, 2009c) also incorporated an extra stage to the linking process, expert review, and this study is also proof that expert review is necessary prior to the project. For instance, in the COPE study, the feedback on the tasks gathered in the very first writing standardisation session lead to changes to the prompts and the stage was repeated. The changes in the prompts also had to be reflected in the test specifications which required revisiting the specification forms of the Manual. Users of the Manual are advised to plan their linking studies accordingly and be prepared to make adjustments as they progress.

7.5.3 For validation frameworks

The research also brought a new perspective to validation models. It pinpointed an area, standard setting, that is deemed significant on its own but has not yet become a fundamental part of validation models, as argued in 7.4.2.

7.6 Areas for further research

As discussed under section 7.3, the study presented in this thesis is a case study, and is thus limited to the experience in a specific context, BUSEL. To evaluate the generalisibility of the findings, the study could be replicated in different contexts. This is important because in other contexts, parameters of aspects of validity that were seen to have been missing in the project reported here might be captured, meaning that the Manual approach to validation has a broader scope than identified in this study or vice versa. Furthermore, different implications of such a linking study on the institutional standards might be determined.

Another area that research can address involves investigating the effectiveness of the validation model suggested in section 7.4.2. The model was developed by the researcher based on the research presented here and was not trialled. Further research might shed light on the practicality of the model and its usability in the type of linking project reported here and in the development of a new examination.

7.7 Concluding remarks

Both the COPE CEFR linking project presented in Chapters 3 and 4 and this research as a whole demonstrated that a CEFR linking project is a huge undertaking. As opposed to the stipulation of the Manual regarding resources, particularly in terms of time, going

through the stages of the process requires a longitudinally orientated framework (this type of project should not be seen as a 'quick fix') and significant human resources. In addition, the process, unlike the Manual's linear presentation, is iterative. Therefore, organisations considering embarking on a linking project, should be well aware of the commitment this endeavour calls for. If they are to undertake such a linking study, they should do their best to ensure that it is carried out meticulously and that it is well resourced.

This study also revealed that a linking project is an invaluable opportunity to establish a comprehensive validation argument for the examinations under study. Institutions endeavouring to undertake such a study should carefully plan in advance so that they can make the most out of this opportunity.

References

Alderson, J.C. (1984). Reading in a foreign language: a reading problem or a language problem? In J.C. Alderson & A. H. Urquhart (Eds.), *Reading in a Foreign Language* (pp.1-24). London: Longman.

Alderson, J. C. (1990a). Testing reading comprehension skills (Part One). *Reading in a Foreign Language, 6*(2), 425-438.

Alderson, J. C. (1990b). Testing reading comprehension skills (Part Two). *Reading in a Foreign Language, 7*(1), 465-503.

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language Test Construction and Evaluation. Cambridge*: Cambridge University Press.

Alderson, J.C. (2000). *Assessing Reading. Cambridge*: CUP.

Alderson, J. C. (2002a). Using the Common European Framework in language teaching and assessment. In J. C. Alderson (Ed.), *Common European Framework of Reference for Languages: learning, teaching, assessment*: *Case studies* (pp. 1-8). Strasbourg: Council of Europe.

Alderson, J. C. (Ed.). (2002b). *Common European Framework of Reference for Languages: Learning, teaching, assessment*: *Case studies.* Strasbourg: Council of Europe.

Alderson, J. C. (2005). *Diagnosing foreign language proficiency: The interface between learning and assessment*. London: Continuum.

Alderson, J. C. (2007). The challenge of (diagnostic) testing: Do we know what we are measuring? In J. Fox & M. Wesche (Eds.), *Language testing reconsidered*: Proceedings of the 27th Language Testing Research Colloquium (LTRC) (pp. 21-39). Ottawa: University of Ottawa Press.

Alderson, J.C. & Bachman, L.F. (2004). *Statistical Analysis for Language Assessment. Cambridge*: Cambridge University Press.

Alderson, J.C. & Huhta, A. (2005). The development of a suite of computer-based diagnostic tests based on the Common European Framework. *Language Testing*. 22 (3), 301-320.

Alderson, J. C., Figueras, N., Kuijper, H., Nold, G., Takala, S., & Tardieu, C. (2004). *The Development of Specifications for Item Development and Classification within the Common European Framework of Reference for Languages: learning, teaching, assessment for Reading and Listening. Final Report of The Dutch CEF Construct Project*. Lancaster: Lancaster University.

Alderson, J.C. & Lukmani, Y. (1989). Cognition and reading: Cognitive levels as embodied in test questions. *Reading in a Foreign Language, 5* (2), 253-270.

Alderson, J. C., Figueras, N., Kuijper, H., Nold, G., Takala, S., & Tardieu, C. (2006).Analysing tests of reading and listening in relation to the Common EuropeanFramework of Reference: The experience of the Dutch CEFR Construct Project. *Language Assessment Quarterly*, 3(1), 3–30.

Allsop, J. (1998). *'Consultancy report'* submitted to the senior management of the Bilkent University School of English Language.

ALTE (1994). The ALTE Code of Practice. Retrieved 07/12/2007, from http://www.alte.org/cop/index.php

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1985). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Anastasi, A. (1954). *Psychological Testing*. New York: MacMillan.

Anderson, G. (1998). *Fundamentals of Educational Research* (2nd Ed.) London: Routledge Falmer.

Anderson, G. (2002). *Fundamentals of Educational Research* (3rd Ed.) London: Routledge Falmer.

Angoff, W. H. (1971). Scales, norms and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement Second Edition* (pp. 508-600). Washington: American Council on Education.

APA (2004). Code of Fair Testing Practices. Retreived on 07/12/2007, from http://www.theaaceonline.com/codefair.pdf

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Bachman, L. (2002). Some reflections on task-based language performance assessment. *Language Testing,* 19(4), 453-476.

Bachman, L.F. (2004). *Statistical Analysis for Language Assessmen*t. Cambridge: CUP.

Bachman, L.F, Davidson, F., Lynch, B., & Milanovic, M. (1996). The use of test method characteristics in the content analysis and design of EFL proficiency tests. *Language Testing*. 13(2), 125-150.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford: Oxford University Press.

Baker, R. (1997). Classical test theory and item response theory in test analysis: Extracts from an investigation of the Rasch model in its application to foreign language proficiency testing. *Language testing update. Special report No. 2*. Lancaster: Lancaster University.

Barr, R., Sadow, M.W. & Blachowicz, C.L.Z. (1990). *Reading diagnosis for teachers: An instructional approach*. New York: Longman.

Bazerman, C. (2000). *Shaping Written Knowledge: The Genre and Activity of the Experimental Article in Science*. Madison, Wisconsin: University of Wisconsin Press.

Bazerman, C. & Prior, P. (2005). Participating in Emergent Socio-literate Worlds: Genre, Disciplinarity, Interdisciplinarity. In R. Beach, J. Green, M. Kamil, & T. Shanahan, (Eds.), *Multidisciplinary Perspectives on Literacy Research Second Edition* (pp. 133-178). Cresskill, NJ: Hampton Press.

Bell, J. (2002). *Doing your research project: a guide for first-time researchers in education, health and social science* (3rd Ed.). Buckingham: Open University Press.

Banerjee, J. (2004). Qualitative analysis methods. Section D of the *Reference Supplement to the preliminary version of the Manual for relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment*. Strasbourg: Council of Europe.

Bereiter, C. & Scardamalia, M. (1987). *The Psychology of Written Composition*. Hillside, NJ: Lawrence Erlbaum Associates.

Bernhardt, E.B. (1991). A psycholinguistic perspective on second language literacy: In J.H. Hulstijn & J.F. Matter (Eds), pp.31-44.

Bingham, W. V. (1937). *Aptitudes and aptitude testing*. New York: Harper & Brothers.

Bloom, B.S. et al. (1956). Taxonomy of educational objectives: The classification of educational goals. *Handbook I: Cognitive domain*. NY: David McKay.

Borg, S. (2004). *Quality in research.* Retrieved 25/11/2008 from http://www.quality- tesol-ed.org.uk/

Borg, S (2006) *Teacher cognition and language education: research and practice.* London: Continuum.

Boorsboom, D., Mellenberg, G.J., & van Heerden, J (2004). The concept of validity. *Psychological Review*, 111, 1061-1071.

Brindley, G. (1991). Defining language ability: The criteria for criteria. In S. Anivan (Ed.), *Current developments in language testing* (pp. 139-164). Singapore: SEAMEO Regional Language Center.

Brindley, G. (2001). Outcomes-based assessment in practice: Some examples and emerging insights. Language Testing, 18, 393-407.

Brown, C., Hedberg, J., & Harper, B. (1994). Metacognition as a basis for learning support software. *Performance Improvement Quarterly,* 7(2), 3-26.

Brown, J.D. (2004). Research methods for applied linguistics: scope, characteristics, and standards. In A. Davis & C. Elder (Eds.), *The Handbook of Applied Linguistics.* Oxford: Blackwell.

Brown, J.D. & Rodgers, T.S. (2002). *Doing Second Language Research*. Oxford: Oxford University Press.

BUSEL (2008). *Bilkent University School of English Language COPE Test Specifications.* Ankara*:* Bilkent University.

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics* 1(1), 1-47.

Canale, M. (1983). From communicative competence to communicative language pedagogy. In J.C. Richards & R. Smidt (Eds) *Language and communication*, London and New York: Longman.

Carmines, E.G. & Zeller, R.A. (1979). *Reliability and Validity Assessment.* Newbury Park, CA: Sage Publications.

Carrell, P. L., & Grabe, W. (2002). Reading. In N. Schmitt (Ed.), *An introduction to applied linguistics* (pp. 233-250). London: Arnold.

Carroll, J.B. (1969). From comprehension to inference. In M.P. Douglas (Ed.). *Thirty-Third Yearbook, Claremont Reading Conference.* Claremont, CA: Claremont Graduate Schools, 39-44.

Carroll, J.B. (1971). Defining language comprehension: Some speculations. *Research Memorandum 71-79*. Princeton, N.J.: Educational Testing Service.

Center for Canadian Language Benchmarks. (2005). *Canadian Language Benchmarks 2000*. Retrieved 10/09/2007, from http://www.cic.gc.ca/english/newcomer/esl-e.html)

Chapelle, C. (1999). Validity in Language Assessment. *Annual Review of Applied Linguistics,*Vol. 9. New York: Cambridge University Press, 254-272.

Cizek, G.J. (1993). Reconsidering standards and criteria. *Journal of Educational Measurement,* 30(2), 93-106.

Cizek, G. J. & Bunch, M. B. (2007). *Standard Setting*. Thousand Oaks, CA: Sage.

Cizek, G.J. (2011). *Reconceptualizing Validity and the Place of Consequences.* Paper Presented at the Annual Meeting of the National Council on Measurement in Education. New Orleans, LA.

Coady, J. (1979). A psycholinguistic model of the ESL reader. In R. Mackay, B. Barkman & R.R. Jordon (Eds), *Reading in a second language* (pp 5-12). Rowleg, MA: Newbury House.

Cobb, T. (2003). VocabProfile, The Compleat Lexical Tutor. Retrieved 13/12/2009, from http://www.lextutor.ca

Cohen, L., Manion, L., & Morrison, K. (2000). *Research methods in education* (5th ed.). London: Routledge.

Cohen, A. D. & Upton, T. A. (2006). Strategies in responding to the new TOEFL reading tasks, *Monograph* 33, Princeton, NJ: Educational Testing Service.

Cohen, A. D. & Upton, T. A. (2007). 'I want to go back to the text': Response strategies on the reading subtest of the new TOEFL. *Language Testing* 24(2), 209-250.

Cohen, L., Manion, L. & Morrison, K. (2007). *Research Methods in Education*. (6th ed.) London: Routledge.

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.

Council of Europe. (2003). *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment*: *Preliminary Pilot Manual.* Strasburg: Council of Europe, Language Policy Division.

Council of Europe. (2004). Reference Supplement to the Preliminary Pilot version of the Manual for *Relating Language examinations to the Common European Framework of Reference for Languages: learning, teaching, assessment.* Strasbourg: Council of Europe, Language Policy Division.

Council of Europe. (2009). *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment*: *Manual.* Strasburg: Council of Europe, Language Policy Division.

Creswell, J. W. (1994). *Research Design: qualitative and quantitative approaches*. Thousand Oaks, California: Sage.

Creswell, J. W. (2009). *Research design: qualitative, quantitative, and mixed methods approaches* (3rd ed.). Los Angeles: Sage.

Cronbach, L.J. (1971). Test Validity. In R.L. Thorndike (Ed.), *Educational Measurement*, 2nd ed. American Council on Education: Washington, DC, 430-507.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52, 281-302.

Cumming, A. (2002). If I had known twelve things. In L. L. Blanton & B. Kroll (Eds), *ESL composition tales: Reflections on teaching* (pp. 123-134). Ann Arbor: University of Michigan Press.

Dawber, T., Lewis, D.M. & Rogers, W.T. (2002). *The cognitive experience of bookmark standard setting participants*, New Orleans, LA: Paper presented at the Annual Meeting of the American Educational Research Association.

Dávid, G.A. (2010). Linking the general English suite of Euro Examinations to the CEFR: a case study report. In W. Martyniuk (Ed.) *Aligning Tests with the CEFR*, *Studies in Language Testing 33* (pp. 177-203). Cambridge: Cambridge University Press.

Davis, F.B. (1944). Fundamental factors of comprehension in reading. *Psychometrica,* 9, 185-197.

Davis, F.B. (1968). Research in comprehension in reading. *Reading Research Quarterly,* 3, 499-545.

Deane, P., Odendahl, N., Quinlan, T., Fowles, M., Bivens-Tatum, J. (2008). *Cognitive models of writing: writing proficiency as a complex integrated skill.* (ETS RR 08-55). Princeton, New Jersey: ETS Retrieved 23/03/2010 from http://www.ets.org/Media/Research/pdf/RR-08-55.pdf

Dechant, Emerald. 1991. *Understanding and teaching reading: An interactive model.* Hillsdale, NJ: Lawrence Erlbaum.

Denscombe, M. (2002) *Ground Rules for Good Research.* Maidenhead: Open University Press.

dos Santos Lonsdale, M. (2005). *Effects of the typographic layout of reading examinations materials on performance*. Unpublished PhD Thesis, University of Reading, UK.

Downey, N. & Kollias, C. (2009). Mapping the Advanced Level Certificate in English (ALCE) examination onto the CEFR. In W. Martyniuk (Ed.) *Aligning Tests with the CEFR*, *Studies in Language Testing 33* (pp. 119-130). Cambridge: Cambridge University Press.

Dörnyei, Z. (2007). *Research Methods in Applied Linguistics*. Oxford: Oxford University Press.

EALTA. (2006). *EALTA Guidelines for Good Practice in language testing and assessment*. Retrieved 05/12/2007, from http://www.ealta.eu.org/documents/archive/guidelines/English.pdf

Eckes, (2009). Many-Facet Rasch Measurement. Section H of the *Reference Supplement to the preliminary version of the Manual for relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching,assessment*. Strasbourg: Council of Europe.

Enright, M. K., Grabe, W., Koda, K., Mosenthal, P., Mulcahy-Ernt, P., & Schedl, M. (2000). *TOEFL 2000 reading framework: A working paper* (TOEFL Monograph No. MS-17). Princeton, NJ: ETS.

ETS (2002). *ETS Standards for Quality and Fairness*. Retreived on 09/12/2007, from http://www.ets.org/Media/About_ETS/pdf/standards.pdf

Faggen, J. (1994). *Setting standards for constructed response tests*: *An overview* (ETS RM- 94- 119). Princeton, NJ: ETS.

Figueras, N. (2007). The CEFR, a lever for the improvement of language professionals in Europe. *Modern Language Journal,* 91, 673-675.

Figueras, N. & Melcion, J. (2002). The Common European Framework in Catalonia. In J. C. Alderson (Ed.), *Common European Framework of Reference for Languages: learning, teaching, assessment*: *Case studies* (pp. 19-24). Strasbourg: Council of Europe.

Figueras, N. & J. E. Noijons (2009). *Linking to the CEFR levels: Research perspectives.*Arnhem, Cito/EALTA.

Figueras, N. & Verhelst, N. (2005) Relating examinations to the Common European Framework: a manual. *Language Testing* 22 (3): 261-279.

Finn, P. (1990). *Helping children learn to read.* White Plains, NY: Longman.

Flesch, R. (1955). *Why Johnny can't read.* New York: Harper and Row.

Fries, C.C. (1963). *Linguistics and Reading*. New York: Holt, Rinehart and Winston.

Frisbie, D. A. (1988). *Reliability of scores from teacher-made tests, Educational Measurement: issues and practice,* 7, 25-35.

Fulcher, G. (2003). *Testing Second Language Speaking*. Edinburgh: Pearson Education Limited.

Fulcher, G. (2004a). *Are Europe's tests being built on an 'unsafe' framework?* Retrieved 20/09/2006, from http://education.guardian.co.uk/tefl/story/0,5500,1170569,00.html

Fulcher, G. (2004b). Deluded by artifices? The Common European Framework and harmonization. *Language Assessment Quarterly*, 1(4), 253-266.

Generalitat de Catalunya. (2006). *Proficiency scales: The Common European Framework of Reference for Languages in the Escoles Oficials d'Idiomes in Catalunya*. Madrid: Cambridge University Press.

Goodman, K. S. 1985. "Unity in reading." In H. Singer & B.R. Robert (Eds) *Theoretical models and the processes of reading.* 3rd edition (pp.813-340). Newark, DE: International Reading Association.

Gough, P. B. (1972). One Second of Reading, *Visible Language 6(4). 291-320.*

Gough, P. B. 1985. "One second of reading." In H. Singer & B.R. Robert (Eds) *Theoretical models and the processes of reading.* 3rd edition (pp.687-688). Newark, DE: International Reading Association.

Gove, M. K. 1983. "Clarifying teacher's beliefs about reading." *The Reading Teacher*. Newark, DE: International Reading Association.

Grabe, W. & Kaplan, R.B. (1996). *Theory and Practice of Writing: An Applied Linguistics Perspective.* London: Longman.

Grabe, W. & Stoller, F. (1997). Reading and vocabulary development in a second language: A case study. In J. Coady & T. Huckin (Eds.), *Second language vocabulary acquisition: A rationale for pedagogy* (pp. 98-122). Cambridge: Cambridge University Press.

Graesser, A.C., McNamara,D.S., Louwerse,M.M, & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, and Computers*, 2004, (36), 193-202.

Green, A. (2009). *The Password Test*. University of Bedforshire. Retrieved 03/07/2010 from http://www.englishlanguagetesting.co.uk/uploads/password-test-design-and-development.pdf

Guilford, J.P. (1946). New standards for test evaluation. *Educational and Psychological Measurement*, 6, 427-439.

Gulliksen, H. (1950). *Theory of mental tests.* New York: Wiley.

Gunthrie J.T. & Kirsch, I.S. (1987). Distinctions between reading comprehension and locating information in text. *Journal of Educational Psychology*, 79, 220-297.

Güven, H., Kantarcıoğlu. E. & Thomas,C. (2009). *Linking Bilkent University School of English language examinations to the CEFR*. Paper presented at Ankara University Foundation Schools 2nd International Conference, May 2009. Ankara.

Hambleton, R.K., Jaeger, R.M., Plake, B.S., & Mills, C. (2000). Setting performance standards on complex educational assessments. *Applied Psychological Measurement*, 24, 355-366.

Hambleton, R. K., & Plake, B. S. (1995). Using an Extended Angoff procedure to set standards on complex performance assessments. *Applied Measurement in Education*, 8(1), 41-55.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of Item Response Theory*. London: Sage Publications.

Hambleton, R.K., Brennan, R.L., Brown, W., Dodd, B., Forsyth, R.A., Mehrens, W.A., Nellhaus, J., Reckase, M.D., Rindone, D. van der Linden, W.J., & Zwick, R. (2000). A response to "Setting Reasonable and Useful Performacne Standards" in the National Academy of Sciences' Grading the Nation's Report Card. *Educational Measurement: Issues and Practice*, 19(2), 5-14.

Hannan, B. & M. Daneman (2001). A new tool for measuring and understanding individual differences in the component processes of reading comprehension. *Journal of Educational Psychology* 93: 103-128.

Hasselgreen, A. (2005). Assessing the language of young learners. *Language Testing*. 22 (3): 337-354.

Hawkey, R. & Barker, F. (2004). Developing a common scale for the assessment of writing. *Assessing Writing*. 9(2).

Hayes, J. R. (1996). A new framework for understanding cognition and affect in writing. In C. Levy & S. Ransdell (Eds.), *The Science of Writing*. NJ: Earlbaum.

Hayes, J. R. & Flower, L. (1980). Identifying the organisation of writing process. In. L. Gregg & E. Steinberg (Eds.). *Cognitive processes in writing* (pp.3-30). N.J: LEA.

Holstein, J. A., & Gubrium, J. F. (2004). The active interview. In D. Silverman (Ed.), *Qualitative research: Theory, method and practice* (pp. 140-161). London: Sage Publications.

Hoover, W. A., & Gough, P. B. (1990). The simple view of reading. *Reading and Writing: An Interdisciplinary Journal*, 2,127-160.

Hoover, W., & Tunmer, W. (1993). The components of reading. In G. Thompson, W. Tunmer, & T. Nicholson (Eds.). *Reading acquisition processes (pp.1-19)*. Clevedon, England: Multilingual Matters.

Hudson, T. (2005). Trends in assessment scales and criterion-referenced language assessment. *Annual Review of Applied Linguistics*, 25, 205-227.

Huhta, A., Luoma, S., Oscarson, M., Sajavaara, K., Takala, S., & Teasdale, A. (2002). DIALANG: A diagnostic language assessment system for adult learners. In J. C. Alderson (Ed.), *Common European Framework of Reference for Languages: learning, teaching, assessment*: *Case studies* (pp. 130-145). Strasbourg: Council of Europe.

Hull, C.L. (1928). *Aptitude testing*. London: Harrap.

Hulstijn, J.H. (2007). The Shaky Ground Beneath the CEFR: Quantitative and Qualitative Dimensions of Language Proficiency. *The Modern Language Journal.* 91(4), 663-667.

Hymes, D. (1972). On communicative competence. In J. Pride & A. Holmes (Eds) *Sociolinguistics* (pp 269-293), Harmsworth and New York; Penguin.

Interagency Language Roundtable. (2011). *An overview of the history of the ILR Language proficiency skill level descriptions and scale by Dr. Martha Herzog.* Retrieved 21/02/2011 from http://www.govtilr.org/Skills/IRL%20Scale%20History.htm

Jaakkola, H., Simonkyla, U.V. & Komsi, K. (2002). How to Promote Learning to Learn in First Foreign Language Classes in Finland. In J. C. Alderson (Ed.), *Common European Framework of Reference for Languages: learning, teaching, assessment*: *Case studies* (pp. 40-52). Strasbourg: Council of Europe.

Jones, N. (2002). Relating the ALTE Framework to the Common European Framework of Reference. In J. C. Alderson (Ed.), *Common European Framework of Reference for Languages: learning, teaching, assessment*: *Case studies* (pp. 167-183). Strasbourg: Council of Europe.

Kaftandjieva, F. & Takala, S. (2002). Council of Europe Scales of Language Proficiency: A validation study. In J. C. Alderson (Ed.), *Common European Framework of Reference for Languages: learning, teaching, assessment*: *Case studies* (pp. 106-129). Strasbourg: Council of Europe.

Kaftandjieva, F. (2004). Standard setting. Section B of the *Reference Supplement to the preliminary version of the Manual for relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment.* Strasbourg: Council of Europe.

Kaiser & Smith 2001 *The Standards for Educational and Psychological Testing: Zugzwang for the Practicing Professional?* Paper presented to The International Personnel Management Association Assessment Coucil. Newport Beach, CA.

Kamberelis, G. (1999). Genre development and learning: Children writing stories, science reports, and poems. *Research in the Teaching of English, 33*(4), 403-460.

Kane, M. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research*, 64(3), 425-461.

Kane, M. (2001). So much remains the same: conception and status of validation in setting standards. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 53-88). Mahwah, N.J.: Lawrence Erlbaum Associates.

Khalifa, H. & ffrench, A. (2008). 'Aligning Cambridge ESOL examinations to the CEFR: Issues & practice' in *IAEA: Re-interpreting Assessment: Society, Measurement and Meaning*. Cambridge: Cambridge Assessment. Retrieved on 20/03/2009 from http://www.iaea2008.cambridgeassessment.org.uk/ca/digitalAssets/180399_Khalifa.pdf

Khalifa, H. & Weir, C. (2009). *Examining reading: Research and practice in assessing second language reading, Studies in Language Testing 29* Cambridge: Cambridge University Press.

Komorowska, H. (2002). The Common European Framework in Poland. In J. C. Alderson (Ed.), *Common European Framework of Reference for Languages: learning, teaching, assessment*: *Case studies* (pp. 9-18). Strasbourg: Council of Europe.

Kunnan, A.J. (1994). Modeling relationships among some test taker characteristics and performance on EFL tests: An approach to construct validation. *Language Testing*, 11 (3), 225-252.

Kunnan, A.J. (1995). *Test taker Characteristics and Test Performance: A Structural Modelling Approach.* Cambridge: University of Cambridge Local Examinations Syndicate and Cambridge University Press.

LaBerge, D. & Samuels, S.J. (1985). "Toward a theory of automatic information processing in reading." In Singer & Ruddell (Eds.) *Theoretical models and the processes of reading.* 3rd edition (pp.689-718). Newark, DE: International Reading Association.

Langenfeld, T. E., & Crocker, L. M. (1994). The evolution of validity theory: Public school testing, the courts, and incompatible interpretations. *Educational Assessment, 2*(2), 149–165.

Lenz, P. & Schneider, G. (2002). Developing the Swiss Model of the European Language Portfolio. In J. C. Alderson (Ed.), *Common European Framework of Reference for Languages: learning, teaching, assessment*: *Case studies* (pp. 68-86). Strasbourg: Council of Europe.

Linacre, J.M. (1989). *Many-facet Rasch measurement*. Chicago: MESA Press.

Linacre, J. M. (1994). *Many-facet Rasch measurement* (2nd ed.). Chicago: MESA Press.

Linacre, J. M. (2007). *FACETS Rasch measurement computer program*. Chicago: Winsteps.com.

Linacre, J. M., & Wright, B. D. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, 8(3), 370.

Liskin-Gasparro, J. (1984). The ACTFL proficiency guidelines: A historical perspective. In T. V. Higgs (Ed.), *Teaching for proficiency: The organizing principle* (pp 11-42). Lincolnwood, IL: National Textbook Co.

Little, D. (2002). Meeting the English Language Needs of Refugees in Ireland. In J. C. Alderson (Ed.), *Common European Framework of Reference for Languages: learning, teaching, assessment*: *Case studies* (pp. 53-67). Strasbourg: Council of Europe.

Little, D. (2002). The European Language Portfolio: Structure, origins, implementation and challenges. *Language Testing*, 35(3), 182-189.

Little, D. (2005). The Common European Framework and the European Language Portfolio: involving learners and their judgments in the assessment process. *Language Testing* 22 (3), 321-336.

Little, D. (2007). The Common European Framework of Reference for Languages: Perspectives on the making of supranational language education policy. *The Modern Language Journal*, 91(4), 645-655.

Lord, F.M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NY: Earlbaum.

Lumley, T. (1993). The notion of subskills in reading comprehension tests: An EAP example. *Language Testing*, 10(3), 211-235.

Martyniuk, W. (2006). *Relating language examinations to the Common European Framework of Reference for Languages (CEFR)*. Paper presented at the Into Europe-European Standards in Language Assessment Conference, Budapest, Hungary. Retrieved 20/9/2006, from http://www.examsreform.hu/Media/Relatinglanguage_exam.ppt.

Martyniuk, W. (Ed) (2010). *Aligning Tests with the CEFR*, *Studies in Language Testing 33*. Cambridge: Cambridge University Press.

Mathews, M. (1990). Skill Taxonomies and Problems for the Testing of Reading. *Reading in a Foreign Language, 7(1), 1990.*

Maxwell, J.A. (1996). *Qualitative Research Design: an interactive approach*. Thousand Oaks, CA: Sage.

McCormick, Thomas W. 1988. *Theories of reading in dialogue: An interdisciplinary study.* New York: University Press of America.

McNamara, T. (1996). *Measuring Second Language Performance*. New York: Addison Wesley Longman.

McNamara, T. (2003). *Validity and reliability in the senior school curriculum: new takes and old questions*. Invited presentation, Australasian Curriculum, Assessment & Certification Authorities (ACACA) 2003 National Conference, Adelaide, July 31. Retrieved on 07.06.2008 from http://www.saceboard.sa.edu.au/acaca/pdf/10professortimmcnamara.pdf

McNamara, D.S., Louwerse, M.M. & Graesser, A.C. (2002). *Coh-Metrix: Automated cohesion and coherence scores to predict text readability and facilitate comprehension*. Technical report, Institute for Intelligent Systems, University of Memphis, Memphis, TN.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 13-103). New York: Macmillan.

Messick S. *(1990).* Validity of Test Interpretation and Use. *Research Report. RR-90-11.* Princeton, NJ*:* Education Testing Service*.*

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons; responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50: 741-9.

Messick, S. (1996). Validity and washback in language testing. *Language Testing*, 13(4), 241-256.

Miles, M. B. & Huberman, A. M. (1994). *Qualitative Data Analysis*. California: Sage Publications.

Mislevy, R.J.; Almond, R.G. & Lukas, J.F. (2004). *A Brief Introduction to Evidence-Centered Design CSE Report 632*. Retrieved on 02.06.2008. from ERIC database.

Mislevy, R.J., Steinberg, Linda S. & Almond, Russell G.(2003) 'Focus Article: On the Structure of Educational Assessments', *Measurement: Interdisciplinary Research & Perspective,* 1: 1, 3-62. Retreived on 08.01.2010. from http://pdfserve.informaworld.com/114045_758064766_785315647.pdf

Moore, T. Morton, J., Price, S. (2010). *Construct validity in the IELTS academic reading test: a comparison of reading requirements in IELTS test items and in university study.* IELTS Research Report. Retrieved on 01/03/2011, from http://www.ielts.org/PDF/vol11_report_4_construct_validity.pdf

Muijs, D. (2004). *Doing quantitative research in education with SPSS.* Thousand Oaks, CA: Sage.

Munby, J.L., (1978). *Communicative Syllabus Design*. Cambridge: Cambridge University Press.

Nakamura, Y. (2001). *Rasch Measurement and Item Banking: Theory and Practice*. Retrieved on 07.07.2009 from ERIC database.

National Center for Research on Evaluation Standards and Student Testing. (1999). CREST Assessment Glosary. Retrieved on 22/10/2008, from www.cse.ucla.edu/CRESST/pages/glossary.htm

Noijons & Kuijper, (2010). Mapping the Dutch foreign language state examinations onto the CEFR. In W. Martyniuk (Ed.) *Aligning Tests with the CEFR*, *Studies in Language Testing 33* (pp.247-265). Cambridge: Cambridge University Press.

North, B. (1997). Perspectives on language proficiency and aspects of competence. *Language Teaching*, 30, 93-100.

North, B. (Ed.) (1992). *Transparency and Coherence in Language Learning in Europe: Objectives, Assessment and Certification*. Symposium held in Ruschlikon, Switzerland, 10–16 November 1991. Strasbourg: Council for Cultural Cooperation.

North, B. (2000). *The development of a common framework scale of language proficiency.* New York: Peter Lang.

North, B. (2002a). Developing Descriptors Scales of Language Proficiency for the CEF Common Reference Levels. In J. C. Alderson (Ed.), *Common European Framework of Reference for Languages: learning, teaching, assessment*: *Case studies* (pp. 87-105). Strasbourg: Council of Europe.

North, B. (2002b). A CEF-Based Self-Assessment Tool for University Entrance. In J. C. Alderson (Ed.), *Common European Framework of Reference for Languages: learning, teaching, assessment*: *Case studies* (pp. 146-166). Strasbourg: Council of Europe.

North, B. (2006). The Common European Framework of Reference: Development, Theoretical and Practical Issues. *IATEFL'S TEA SIG Newsletter* Summer 2006 (11-35).

North, B. (2008) The CEFR levels and descriptor scales. In L. Taylor & C. Weir (Eds), 21–66. *Studies in Language Testing 27; Multilingualism and Assessment – Achieving transparency, assuring quality, Sustaining diversity – Proceedings of the ALTE Berlin Conference, May 2005.* Cambridge: Cambridge University Press.

North, B. & Scheinder, G. (1998). Scaling descriptors for language proficiency scales. *Language Testing* 15 (2): 217-263.

O'Dwyer, J. (2008). *Formative Evaluation for Organizational Learning*. Frankfurt: Peter Lang.

Oppenheim, A.N. (1992). *Questionnaire Design, Interviewing and Attitude Measurement.* London: Pinter.

O'Sullivan, B. (2000). *Towards a Model of Performance in Oral Language Testing*. Unpublished PhD Dissertation. University of Reading.

O'Sullivan, B. 2006. Testing Language for Business: a critical overview of current practice. *ESP Malaysia*, 3: 17-31.

O'Sullivan, B. (2008). *Modelling Performance in Tests of Spoken Language*. Frankfurt: Peter Lang.

O'Sullivan, B. (2009a). *City & Guilds Communicator Level IESOL Examination (B2) CEFR Linking Project Case Study Report*. City & Guilds Research Report. Retrieved on 29/1/2009, from http://www.cityandguilds.com/documents/ind_general_learning_esol/CG_Communicator_Report_BOS.pdf

O'Sullivan, B. (2009b). *City & Guilds Achiever Level IESOL Examination (B1) CEFR Linking Project Case Study Report*. City & Guilds Research Report.

O'Sullivan, B. (2009c). *City & Guilds Expert Level IESOL Examination (C1) CEFR Linking Project Case Study Report*. City & Guilds Research Report.

O'Sullivan, B. (2010). The City & Guilds Communicator examination linking project: a brief overview with reflections on the process. In W. Martyniuk (Ed.) *Aligning Tests with the CEFR*, *Studies in Language Testing 33* (pp.33-49). Cambridge: Cambridge University Press.

O'Sullivan, B. (2011). Language Testing. In J. Simpson (Ed.). Routledge. *Handbook of Applied Linguistics* (pp.259-273). Oxford: Routledge.

O'Sullivan & Weir, C.J. (2011) Language testing = validation. In B. O'Sullivan (Ed.) *Language Testing: Theories and Practices* (pp.13-32). Basingstoke: Palgrave 'Advances in Linguistics' series.

Pant, H.A., Rupp, A.A., Tiffin-Richards, S.P. & Köller, O. (2009), "Validity Issues in Standard-Setting Studies", *Studies in Educational Evaluation*, Vol. 35, pp. 95–101.

Papageorgiou, S. (2007a). *Relating the Trinity College London GESE and ISE exams to the Common European Framework of Reference: Piloting of the Council of Europe draft Manual. (Final project report).* Lancaster: Lancaster University.

Papageorgiou, S. (2007b). *Setting standards in Europe: The judges' contribution to relating language examinations to the Common European Framework of Reference*. Unpublished PhD dissertation. University of Lancaster.

Pearson Graduate Management Admission Council (2012). Pearson Test of English academic. Retrieved on 24/03/2012, from http://ganeshonline.net/Documents/USScoreIntepretation.pdf

Perfetti, C.A. (1977). Language comprehension and fast decoding: some psycholinguistic prerequisites for skilled reading comprehension (pp 141-183). In J.T. Guthrie (Ed.) *Cognition, curriculum, and comprehension* (pp. 20–41). Newark, DE: International Reading Association.

Perfetti, C.A. (1985). *Reading Ability*. New York: Oxford University Press.

Pizorn, K. (2009). Designing proficiency levels for English for primary and secondary school students and the impact of the CEFR. In N. Figueras & J. E. Noijons (Eds). *Linking to the CEFR levels: Research perspectives* (pp.87-102). Arnhem, Cito/EALTA.

Popham, W.J. (2005). As Always, Provocative. *Journal of Educational Measurement,* 15(4), 297-300.

Purpura, J. (1999). *Modeling the relationships between test takers' reported cognitive and metacognitive strategy use and performance in language tests.* Cambridge: University of Cambridge Local Examinations Syndicate and Cambridge University Press.

Qian, D.D. (2002). Investigating the relationship between vocabulary knowledge and academic reading performance: An assessment perspective. *Language Learning*, 52: 513-536.

Qian, D.D. & M. Schedl, 2004. Evaluation of an in-depth vocabulary knowledge measure for assessing reading performance. *Language Testing* 21(1), 28-52.

Robson, C. (1993). Real *World Research: A Resource for Social Scientists and Practitioner Researchers*. Oxford: Blackwell.

Robson, C. (2002). Real World Research. Malden: Blackwell Publishing. (2nd edition)

Roehampton University & Universidad Veracruzana. (2008). *Exaver: affordable language test development project.* Retrieved on 8/6/2008, from http://www.uv.mx/exaver/nuv/index.html.

Rosenshine, B.V. (1980). Skill hierarchies in reading comprehension. In Spiro, R.J. et al. (Eds.), pp. 535-554.

Ruddell, R.B., & Speaker, R. (1985). The interactive reading process: A model. In Singer, H. & Ruddell, B.R. *Theoretical models and the processes of reading.* 3rd edition (751-793). Newark, DE: International Reading Association.

Rumelhart, D. E. (1985). Toward an interactive model of reading. In Singer, H. & Ruddell, B.R. *Theoretical models and the processes of reading.* 3rd edition (pp. 722-750). Newark, DE: International Reading Association.

Schraw, G., Wade, S. E., & Kardash, C. A. M. (1993). Interactive effects of text-based and task-based importance on learning from text. *Journal of Educational Psychology,* 85. Pp. 652–661.

Schedl, Gordon, Carey & Tang, (1996). Proficiency descriptors based on a scale-anchoring study of the new TOEFL iBT reading test. *Language Testing*, 24 (3) 417–444.

Scott, M. (2009). WordSmith Tools 5.0. Oxford University Press.

Shaw, S. & Weir, C. J. (2007). *Examining writing: Research and practice in assessing second language writing, Studies in Language Testing 26.* Cambridge: Cambridge University Press and Cambridge ESOL.

Shepard, L.A. (1993). Evaluating test validity. In Darling-Hammond, L. (Ed.) *Review of Research in Education,* Vol. 19, 405-450. Washington, DC: American Educational Research Association.

Shepard, L.A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice,* 16, 5-8.

Silverman, D. (1993). *Interpreting qualitative data: Methods for analyzing talk, text and interaction*. London: Sage.

Silverman, D. (2000). *Doing Qualitative Research: A Practical Handbook*. London: Sage.

SIOP (2003). *Principles for the Validation and Use of Personnel Selection Procedures*. Retrived 12/07/2008, from http://www.siop.org/_Principles/principles.pdf

Stake, R.E. (1995). The Art of Case Study Research. Thousand Oaks, CA: Sage.

Stanovich, K. E. (1980). Toward an Interactive Compensatory Model of Individual Differences in the Development of Reading Fluency, *Reading Research Quarterly* 16(1), 32-71.

Stanovich, K. E. (1984). The Interactive-compensatory model of reading: A confluence of developmental, experimental and educational psychology. *Remedial and Special Education,* 5, 11-19.

Takala, S. (Ed.) (2004). *Reference Supplement to the preliminary version of the Manual for relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching and assessment*. Strasbourg: Council of Europe.

Tannenbaum, R. J. & Wylie, E. C. (2005). *Mapping English Language Proficiency Test Scores onto the Common European Framework.* Research Report 05-18. Princeton, NJ: Education Testing Services.

Tannenbaum, R. J. & Wylie, E. C. (2007). *Mapping TOEFL iBT, TOEIC, and TOEIC Bridge on to the Common European Framework: Interim Report.* Princeton, NJ: Education Testing Services.

Tannenbaum, R. J. & Wylie, E. C. (2008). *Linking English Language Test Scores o to the Common European Framework of Reference: An Application of Standard Setting Methodology.* Princeton, NJ: Education Testing Services.

Taylor, L (2004) IELTS, Cambridge ESOL examinations and the Common European Framework, *Research Notes* 18, 2–3.

Taylor, L., & Jones, N. (2006). Cambridge ESOL exams and the Common European Framework. Cambridge ESOL *Research Notes*, May 2006(24), 2-5. Retrieved 10/10/2007, from http://www.cambridgeesol.org/rs_notes/rs_nts2024.pdf.

Thomas, C. (2009). *From tension to synergy – the Bilkent experience*. Paper presented at the sixth Annual Conference of EALTA, Turku.

Thomas, C. & Kantarcıoğlu, E. (forthcoming). *COPE CEFR linking project report*. Ankara: Bilkent University.

Thomas, R.M. 2003. *Blending Qualitative and Quantitative Research Methods in Theses and Dissertations*. California: Corwin Press Inc.

Thorndike, R. L., & Hagen, E.P. (1959). *Ten Thousand Careers*. New York: John Wiley & Sons, Inc.

Thorndike, R.L. (1971). *Reading as reasoning*. Address delivered to Division 15, American Psychological Association, Washington DC.

Thorndike, R.L. (1973). Reading comprehension, education in 15 countries: An empirical study. *International Studies in Education, 3.*

Traub & Rowley, (1991). Understanding Reliability. *Educational Measurement: Issues and Practice*, 19(1), 37-45.

Trinity College. Undated. *Relating the Trinity College London GESE and ISE examinations to the Common European Framework of Reference – project summary*. London: Trinity College London.

Urquhart, A.H. (1987). Comprehensions and Interpretations. *Reading in a Foreign Language,* 3(2), 387-410.

Urquhart, A.H. & Weir, C. J. (1998). *Reading in a Second Language: Process, Product and Practice*. London: Longman.

Vandergrift, L. (2006). *Proposal for a Common Framework of Reference for Languages for Canada*. Social Sciences and Humanities Research Council of Canada Heritage.

Verhelst, N. (2004a). Classical Test Theory. Section C of the *Reference Supplement to the preliminary version of the Manual for relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment*. Strasbourg: Council of Europe.

Verhelst, N. (2004b). Factor Analysis. *Section F of the Reference Supplement to the preliminary version of the Manual for relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment*. Strasbourg: Council of Europe.

Verhelst, N. (2004c). Generalizability Theory. *Section E of the Reference Supplement to the preliminary version of the Manual for relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment*. Strasbourg: Council of Europe.

Verhelst, N. (2004d). *Item Response Theory. Section G of the Reference Supplement to the preliminary version of the Manual for relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment*. Strasbourg: Council of Europe.

Weaver, C. (1990). *Understanding whole language: From principles to practice.* Portsmouth, NH: Heinemann.

Weigle, S.C. (2002). *Assessing Writing*. Cambridge: CUP.

Weir, C.J. (1994). *Reading as a multi-divisible or unitary: between Scylla and Charybdis.* Paper presented at the RELC, SAEMEO Regional Language Centre, Singapore.

Weir, C.J. (1993). *Understanding and Developing Language Tests*. London: Prentice Hall.

Weir, C. J. (2005a) *Language Testing and Validation An Evidence-Based Approach*. Oxford: Palgrave.

Weir, C. J. (2005b). Limitations of the Common European Framework for developing comparable examinations and tests. *Language Testing*, 22(3), 281–300.

Weir, C.J. & Khalifa, H. (2008a). 'A cognitive processing approach towards defining reading comprehension'. *Research Notes, Cambridge ESOL* 31: 2-10

Weir, C.J. & Khalifa, H. (2008b). 'Applying a cognitive processing model to Main Suite Reading papers'. *Research Notes, Cambridge ESOL* 31: 11-16

Weir, C. J., & Porter, D. (1994). The Multi-Divisible or Unitary Nature of Reading: The language tester between Scylla and Charybdis. *Reading in a Foreign Language*, 10(2), 1-19.

Weir, C.J., Devi, S., Green, A.B., Hawkey, R., Maniski, T. Unaldi, A. & Zegarac, V. (2006). *The relationship between the academc reading construct as measured by IELTS and the reading experiences of stduents in the first year of their courses at a British University*. IELTS Research Report. Retrieved 01/03/2011 from http://www.ielts.org/PDF/Vol9_Report4.pdf

Weir, C.J., Green, A.B., Hawkey, R. & Unaldi, A. (2008). *The cognitive process underlying the academic reading construct as measured by IELTS*. IELTS Research Report. Retrieved 01/03/2011 from http://www.ielts.org/PDF/Vol9_Report4.pdf

Wertenschlag, L., Muller, M. & Schmitz, H. (2002). The Common European Framework and the European Level Descriptions for German as a Foreign Language. In J. C. Alderson (Ed.), *Common European Framework of Reference for Languages: learning, teaching, assessment*: *Case studies* (pp. 184-197). Strasbourg: Council of Europe.

Wilson, K.M., (1999). *Validity of Global Self-Ratingsof ESL Speaking Proficiency Based on an FSI/ILR-Referenced Scale.* Research Report 13. ETS New Jersey, Princeton. Retreived 30.12.2010. from http://www.ets.org/Media/Research/pdf/RR-99-13.pdf

Wright, B. D. & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, 8(3), 370.

Wu, R. (2011). *Establishing the validity of the General English Proficiency Test reading component through a critical evaluation on alignment with the Common European Framework of Reference*. Unpublished PhD Dissertation. University of Bedforshire.

Yin, R.K. (2009). *Case Study Research Design and Methods* 4th Edition. Thousand Oaks, CA: SAGE Publications.

Zieky, M.J. & Perie, M. (2006). *A primer on setting cutscores on tests of educational achievement*. Retrieved 25/01/2008, from http://www.etseuropepi.org/fileadmin/free_resources/Institutional_website/Policy_and_Research/Cut_Scores_Primer.pdf

Zeiky, M.J., Perie, M., & Livingston, S. (2008). *Cutscores: A Manual for Setting Standards of Performance on Educational and Occupations Tests.* Princeton, NJ: Educational Testing Service.

AN OVERVIEW

Benchmark	Proficiency Level	Speaking and Listening Competencies	Reading Competencies	Writing Competencies
STAGE I: BASIC PROFICIENCY				
1 2 3 4	Initial Developing Adequate Fluent	Creating/interpreting oral discourse in routine non-demanding contexts of language use in: • Social interaction • Instructions • Suasion (getting things done) • Information	Interpreting simple texts: • Social interaction texts • Instructions • Business/service texts • Informational texts	Creating simple texts: • Social interaction • Recording information • Business/service messages • Presenting information
STAGE II: INTERMEDIATE PROFICIENCY				
5 6 7 8	Initial Developing Adequate Fluent	Creating /interpreting oral discourse in moderately demanding contexts of language use in: • Social interaction • Instructions • Suasion (getting things done) • Information	Interpreting moderately complex texts: • Social interaction texts • Instructions • Business/service texts • Informational texts	Creating moderately complex texts: • Social interaction • Reproducing information • Business/service messages • Presenting information/ ideas
STAGE III: ADVANCED PROFICIENCY				
9 10 11 12	Initial Developing Adequate Fluent	Creating/interpreting oral discourse in very demanding contexts of language use in: • Social interaction • Instructions • Suasion (getting things done) • Information	Interpreting complex and very complex texts: • Social Interaction texts • Instructions • Business/service texts • Informational texts	Creating complex and very complex texts: • Social interaction • Reproducing information • Business/service messages • Presenting information/ ideas

APPENDIX 2B A summary of the ILR scales

0	No proficiency
0+	Memorized proficiency
1	Elementary proficiency
1+	Elementary proficiency, Plus
2	Limited working proficiency
2+	Limited working proficiency, Plus
3	General professional proficiency
3+	General professional proficiency, Plus
4	Advanced professional proficiency
4+	Advanced professional proficiency, Plus
5	Functionally native proficiency

APPENDIX 2C The link between ILR scales and ACTFL rating scales

ILR Scale	ACTFL Scale	Definition
5	Native	Able to speak like an educated native speaker
4+ 4	Distinguished	Able to speak with a great deal of fluency, grammatical accuracy, precision of vocabulary and idiomaticity
3+ 3	Superior	Able to speak the language with sufficient structural accuracy and vocabulary to participate effectively in most formal and informal conversations
2+	Advanced Plus	Able to satisfy most work requirements and show some ability to communicate on concrete topics
2	Advanced	Able to satisfy routine social demands and limited work requirements
1+	Intermediate - High	Able to satisfy most survival needs and limited social demands
1	Intermediate - Mid Intermediate - Low	Able to satisfy some survival needs and some limited social demands Able to satisfy basic survival needs and minimum courtesy requirements
0+	Novice - High	Able to satisfy immediate needs with learned utterances
0	Novice - Mid Novice - Low 0	Able to operate in only a very limited capacity Unable to function in the spoken language No ability whatsoever in the language

APPENDİX 2D CEFR Global Scale

Proficient User	C2	Can understand with ease virtually everything heard or read. Can summarise information from different spoken and written sources, reconstructing arguments and accounts in a coherent presentation. Can express him/herself spontaneously, very fluently and precisely, differentiating finer shades of meaning even in more complex situations.
	C1	Can understand a wide range of demanding, longer texts, and recognise implicit meaning. Can express him/herself fluently and spontaneously without much obvious searching for expressions. Can use language flexibly and effectively for social, academic and professional purposes. Can produce clear, well-structured, detailed text on complex subjects, showing controlled use of organisational patterns, connectors and cohesive devices.
Independent User	B2	Can understand the main ideas of complex text on both concrete and abstract topics, including technical discussions in his/her field of specialisation. Can interact with a degree of fluency and spontaneity that makes regular interaction with native speakers quite possible without strain for either party. Can produce clear, detailed text on a wide range of subjects and explain a viewpoint on a topical issue giving the advantages and disadvantages of various options.
	B1	Can understand the main points of clear standard input on familiar matters regularly encountered in work, school, leisure, etc. Can deal with most situations likely to arise whilst travelling in an area where the language is spoken. Can produce simple connected text on topics which are familiar or of personal interest. Can describe experiences and events, dreams, hopes and ambitions and briefly give reasons and explanations for opinions and plans.
Basic User	A2	Can understand sentences and frequently used expressions related to areas of most immediate relevance (e.g. very basic personal and family information, shopping, local geography, employment). Can communicate in simple and routine tasks requiring a simple and direct exchange of information on familiar and routine matters. Can describe in simple terms aspects of his/her background, immediate environment and matters in areas of immediate need.
	A1	Can understand and use familiar everyday expressions and very basic phrases aimed at the satisfaction of needs of a concrete type. Can introduce him/herself and others and can ask and answer questions about personal details such as where he/she lives, people he/she knows and things he/she has. Can interact in a simple way provided the other person talks slowly and clearly and is prepared to help.

APPENDIX 3A Familiarisation stage questionnaire

CEF PROJECT FAMILIARISATION STAGE SESSION 1 QUESTIONNAIRE

I. Background Information

1. Years of experience as an English language teacher (Circle ONE)

0-4	5-9	10-14	15-19	20-24	25+

2. Role in BUSEL - Please tick one of the below as appropriate

Teacher ☐
Head of Teaching Unit ☐
Textbook Development Group Member ☐
Curriculum and Testing Unit Member ☐
Member of the Directorate ☐
Other (Please specify) _____

3. How familiar were you with the CEF levels before you took part in the project? (Circle ONE)

Not familiar at all				Very familiar
1	2	3	4	5

4. Read the following statements and decide whether they are true or false. Put a tick under the right column. As you answer the questions, please keep in mind that all the statements are about your knowledge about the CEFR before you were involved in the project.

Before I was invited to take part in the project, I knew that	**FALSE**	**TRUE**
1. the CEFR is a document for language learning, teaching and assessment in Europe.		
2. the CEFR consists of 6 levels.		
3. exams such as FCE, IELTS or TOEFL are linked to the CEFR.		

II. Pre-session Reading Tasks

Circle ONE number for each of the statement below to give your opinion about the pre-session reading tasks.

	Disagree Strongly	Disagree	Not sure	Agree	Agree Strongly
1. The amount of the reading material was manageable.	1	2	3	4	5
2. The amount of time given for the reading was sufficient.	1	2	3	4	5
3. The aim of the tasks was clear.	1	2	3	4	5
4. The tasks had clear instructions.	1	2	3	4	5
5. The task sheets were well-designed.	1	2	3	4	5
6. The content of the reading material was easy to comprehend.	1	2	3	4	5

7. Doing the tasks helped me see how the CEF levels progressed.	1	2	3	4	5

III. Familiarization Session 1

Circle ONE number for each of the statements below to give your opinion about the familiarization session. (For questions 7 to 10, the tasks used in the session are explained at the end of the table.)

	Disagree Strongly	Disagree	Not sure	Agree	Agree Strongly
1. The purpose of the session was clear.	1	2	3	4	5
2. The opening speech about the CEF project gave me a clear idea about the project.	1	2	3	4	5
3. The opening speech about the CEF project gave me a clear idea of my role in the project.	1	2	3	4	5
4. The session was well-designed.	1	2	3	4	5
5. There was a clear progression of tasks in the session.	1	2	3	4	5
6. The pre-session reading tasks were linked to the tasks in the session itself.	1	2	3	4	5
7. Task 1, 2 and 3 followed by a discussion helped clarify the CEF levels better.	1	2	3	4	5
8. Task 4 helped me better understand the CEF levels.	1	2	3	4	5
9. At the end of task 5 and 6, I began to see the differences between the CEF levels.	1	2	3	4	5
10. The wrap-up discussion	1	2	3	4	5
11. The comments and questions I sent prior to the session were addressed.	1	2	3	4	5
12. The individual feedback sheets helped me see my strengths and weaknesses.	1	2	3	4	5
13. The feedback given to the group as a whole was	1	2	3	4	5
14. At the end of session, I felt more confident about	1	2	3	4	5

The tasks carried out in the session are as follows:

Task 1: Identifying key characteristics of the levels in the CEF global scale
Task 2: Analyzing section 3.6 – salient features of CEF levels
Task 3: Poster completion
Task 4: Relating BUSEL levels to CEF levels
Task 5: Reordering CEF global scales followed by feedback and discussion
Task 6: Reordering CEF scales for each skill followed by feedback and discussion

IV. Content

Circle ONE number for each of the statements below to give your opinion about the content of familiarization session 1.

	Disagree Strongly	Disagree	Not sure	Agree	Agree Strongly
1. There is a clear progression between the CEF global scales (CEF Table 1).	1	2	3	4	5
2. There is a clear progression between the CEF self-assessment scales for each skill (CEF Table 2).	1	2	3	4	5
3. The number of levels in CEF is adequate to show language progression.	1	2	3	4	5
4. CEF levels can be used in every context.	1	2	3	4	5
5. The can-do statements used to describe the levels are unambiguous.	1	2	3	4	5
6. I can easily relate the CEF levels to our context.	1	2	3	4	5
7. There is a clear link between CEF levels and the levels commonly used by commercial books i.e. Elementary, Intermediate, Advanced etc.	1	2	3	4	5
8. Aspects of language use are clearly explained in Chapter 4 of the CEF booklet.	1	2	3	4	5
9. Language competences are clearly explained in Chapter 5 of the CEF booklet.	1	2	3	4	5
10. I can now understand the rationale behind the CEF descriptor.	1	2	3	4	5

V. Future needs / demands

As you know, familiarization sessions will continue. Please write down any suggestions or things you would like to see done/covered in the follow-up sessions.

_____

_____

_____

APPENDIX 3B Familiarisation stage field notes coding scheme

Themes	Descriptions	Examples
Context validity	Task design	From B2 onwards context, the target situation is not important anymore The difference here is the topic is reasonably familiar We need to state exactly what the levels are
	Task demands	Mentions background so the expectation is past tense This student should not have any difficulty in understanding any kind of spoken language
	Language knowledge	Formulaic language is required at A2
Cognitive validity	Language knowledge	Grammar items, language functions define a level High command of the language is required to have a deeper understanding
Scoring validity	Criteria	The descriptors are written from an action statement but what is missing is what tells us so There is a problem with the use of 'complex' what exactly is meant? These are intentions not descriptions

APPENDIX 3E Specification stage coding scheme

I: Interviewee

Themes	Descriptions	Skill	Examples
Role of the specification stage	Reflection on the exam	R/W	I2: …helps you look at the exam in depth
	Areas to improve	R/W	I1: …shows areas where you may need to improve of things or things that you haven't perhaps considered before
Test taker	Physical/ physiological	R/W	No cases of this code in the data
	Psychological	R/W	No cases of this code in the data
	Experiential	R/W	No cases of this code in the data
Context validity	Task design	R	I1: You're not actually asked to think a lot about I mean, for the reading comprehension form, form A10. There is not a great deal to think about, what themes are they supposed to deal with? Ok. Which communicative tasks and activities are they expected to? I mean these questions were all fairly straightforward to answer. I don't know. I know I'm rambling on a bit. There is only literally 5 points to consider so there wasn't a lot of actual, there wasn't a lot in there that made me to actually stop and think oh why do we have parts 1 and 2 and 3. SO filling in those forms didn't make me think that.
	Task demands	W	I2: I think for me the real exposer of the problems and the issues the underlying issues regarding our writing prompts for example happened during the discussion stage when we were actually got past that the specification and actually started discussing it
	Test administration	R/W	I2: But I just thought that the administration side of it was not emphasized and I'm not sure if that was because they figure it something that's not relevant to the specifications but I think that you can't really separate administration from the exam if you have you know improper administration, then you're gonna have improper results no matter how good your grading is or how well you've designed your exam
Cognitive validity	Task design	R/W	I2: Whether it's the design stage especially what you were talking about earlier Efser was it a cognitive requirement? When you look at it how much cognitive you know process so until you know a bunch of people look at something and start not just by through filling in the forms but actually through the actual discussion start you know looking at it very critically, it doesn't contribute much I think. Now there might be

			certain areas and there are there are certain areas in these specification forms you know which categorize things and helps you think more categorically and helps you
	Cognitive requirements	R/W	I1: Forms for reception for linguistic competence etc. that actually was the area where made me think about our exam rather than the general one for reading like what is the range of lexical and grammatical competence. That really did make me think and when I look at what we filled in on the forms, it doesn't really make sense. I'd we filled I can't actually remember what we filled in but my note is actually here it doesn't make sense it's just like a list of grammatical structures and things from the CEFR book which is sort of meaningless. I think to fill that properly again what you said you'd have to spend a lot of time really analyzing the lexis, the structures of the actual I'm talking about reading tests. Those were the most challenging and I think thought-provoking. For reading, we had linguistic competence that was valid but sociolinguistic competence one didn't seem to relate to our reading test as far as I can remember. The pragmatic competence and the strategic competence how can you answer that? That's quite a difficult one to think about what strategies we are expecting the test takers to use when they deal with the COPE reading paper. So those I find those really useful. These are the forms that took the longest to fill in.
Scoring validity	Rationale	R/W	I2: It's helpful in that I mean when you look at the grades you know the reporting the results, the rationale
	Criteria	R/W	I1: We don't do enough I think with the COPE exam to help our students. They just get the score means nothing to them. A, B, C or F
	Answer key	R/W	No cases of this code in the data
	Statistical analysis	R/W	You should care but you don't about the statistical data and how it's used and how it's reported and you know as it says in relation to other tests they might take and their future performance you know things like that and I think that on other times so it's useful to have the specification mention those things so those things are also taken into consideration so it doesn't just stop after you give the student a certain grade
Consequential validity (Score	Reporting grades	R/W	Particularly the form reporting results. And again that just threw up an area that we're weak you know how the results are reported.

interpretation)			We don't do enough I think with the COPE exam to help our students
	Tests takers	R/W	I1: There is no information at all given to the candidates on how to interpret those results at all so that was a weak area that I think we definitely need to work on. For reception and it's quite difficult to do it for reception for certainly for writing we could do a lot more to help students we could back up the writing grade with come details, can-do statements. That was useful filling that in cause that did make me think about if we could be doing much more with the results
	Use of grades	R/W	I2: So it's helpful to have the specification take us beyond the marking and scoring and into the data and the data interpretation stages so I thought that's a good thing about it.
	Washback	R/W	I2: it's useful to have the specification mention those things so those things are also taken into consideration so it doesn't just stop after you give the student a certain grade. You know what you do with those grades? And how that data is used afterwards is equally as important
Criterion-related validity (Score value)	Future performance	R/W	I1: There is nothing in here that make you think about you know future performance I2: right in the specification so maybe that's a shortcoming of the specifications. It doesn't really talk much about that. But the fact that it even takes you to you know reporting the results and the rationale for that makes me think about what happens with these scores afterwards. But it could be a shortcoming. There could be a bit added on to it I3: yeah I2: in terms of what is actually done with the score in terms future test performance I1: that's an important point. Yeah that's right. Something should be included.
	Comparison with other tests	R/W	I1: For example with the COPE exam in relation to IELTS, in relation to FCE. What does it mean? What does our score mean or future performance. Can we predict how they're gonna perform in their faculties? FAE you know for example. Things like that.
Level of the exam	Intended level	R/W	I1: I thought it was quite reassuring doing the specifications, looking at the exam, filling in these forms because our gut intuitive feeling backed up by filling in these. It did come out

			as roughly B2 and that was what the group decided as well. We have that graphic profile most of the exam is B2 but it's slightly uneven when it comes to things like I can't remember something like socio whatever I2: linguistic competence I1: that's much lower because in our context they don't need that. It doesn't need to be so high and our what was the other one was C1 the linguistic I3: competence? I1: it was higher as well for reception and it backed up our feeling about the exam. We weren't completely off. So then I though that yeah the standard we set and what we think of the standard is met by filling in these forms
	Understanding of the level	R/W	P1: I think it helps you develop your own understanding of the exam P3: And sometimes you understand that you don't know enough about the exam
	Increasing the standards	R/W	I1: I don't think it would. I think it's going back to what Hakan says it's only when you're actually sitting down, discussing and threshing out the level with a group of people where you can see the you can see what the level really is you can see whether it's too low too high. I2: Maybe sorry to cut you off. Maybe you know like I was saying earlier the categorization that is set in the specification stage. That's helpful you know looking at your exam ok in terms of language then you look at your exam in terms of the strategies the students use you know monitoring repair all that kind of stuff so perhaps referring to the specification and comparing your task, your test to some of those categories then saying well ok we're ok in terms of language but how else I3: Aha. (agreeing) I2: can we modify the difficulty level of our exam by looking at these different categories? How about sociolinguistic competence? Is that relevant for us? No it isn't but is there a category that is relevant for us and how can we increase the challenge

			I1: hmm. I2: or decrease the challenge based on these different categories? I think that would be helpful because I do agree, obviously I agree cause I was saying it earlier as well, that we need to have a it is only through discussion that the development comes through but the discussion also needs to be focused and one way to focus it is like using those categories and the subheadings implied and saying ok what about you know challenge that we ask compared to these compared to these categories? Are we doing enough here or are we doing too much here? And that would help you modify your exam level in a more systematic way.

APPENDIX 3G Standardisation stage reading and writing questionnaires

NAME: _____

EVALUATION OF COPE STANDARD SETTING – reading

The questionnaire below is comprised of 4 sections, viz. Standard Setting Session, Cambridge ESOL and Finnish Matriculation Exam Calibrated Samples, CEF Descriptors and CEF Linking Process, which correspond to aspects of the process undertaken as part of the reading standard setting workshop.

I would be grateful if you would indicate your level of agreement with each of the statements given and add any additional general comments you might have, or any specific comments about any statement in the questionnaire. You can use the back of these sheets, if necessary, to enlarge upon your comments. The information you provide will remain confidential.

I. Standard Setting Session

Item	Statement	Strongly Agree	Agree	Disagree	Strongly Disagree
1	Introductory talk provided me with a clear understanding of the purpose of the session.				
2	The tasks were clearly explained.				
3	The training and practice exercises helped me understand how to perform the tasks.				
4	Discussions aided my understanding of the levels.				
5	Discussions aided my understanding of how performances are assessed.				
6	There was adequate time provided for doing the tasks.				
7	There was adequate time provided for discussions.				
8	There was an equal opportunity for everyone to contribute his/her ideas and opinions.				
9	I was able to follow the instructions accurately.				
10	I was able to complete the judgment sheets accurately.				
11	The discussions after the first round of ratings were helpful to me.				
12	The discussions after the second round of ratings were helpful to me.				
13	The item statistics provided helped me with my judgments about the items.				
14	I am confident about the defensibility of the final recommended cut score.				
15	I am confident about the appropriateness of the final recommended cut score.				
16	There was a productive and efficient working environment.				

Adapted from: Cizek, Bunch, & Koons (2004)

II. Cambridge ESOL and Finnish Matriculation Exam Calibrated Samples

Item	Statements	Strongly Agree	Agree	Disagree	Strongly Disagree
1	The *FCE* samples helped me to carry out the standardization task.				
2	The *CAE* samples helped me to carry out the standardization task.				
3	The *Finnish* samples helped me to carry out the standardization task.				
4	The *FCE* samples were representative of the levels claimed by the exam providers.				
5	The *CAE* samples were representative of the claimed by the exam providers.				
6	The *Finnish* samples were representative of the claimed by the exam providers.				
7	The *FCE* samples were relevant for use in academic contexts.				
8	The *CAE* samples were relevant for use in academic contexts.				
9	The *Finnish* samples were relevant for use in academic contexts.				

III. CEF Descriptors

Item	Statements	Strongly Agree	Agree	Disagree	Strongly Disagree
1	I am satisfied with the definition of "the least able B2 candidate" used in the standard setting.				
2	The actual use of the CEF reading descriptors showed me that they are applicable to *academic* context.				
3	The actual use of the CEF reading descriptors showed me that they are applicable to *any* context.				
4	The CEF reading descriptors catered for all aspects of the skill of reading.				

IV. CEF Linking Process

Item	Statements	Strongly Agree	Agree	Disagree	Strongly Disagree
1	The process helps BUSEL to reconsider the level of the COPE exam.				
2	The process is beneficial for BUSEL in seeing the actual level of the papers in the COPE exam.				
3	The results of the linking process reveal that BUSEL should consider revising the COPE exam.				
4	The linking process makes COPE a more reliable exam.				
5	The linking process contributes to the validation of the COPE exam.				

Comments:

Thank you very much for the time you have put into answering these questions. Your answers will be collated and aggregated for use as part of my PhD research.

NAME: _____

EVALUATION OF COPE STANDARD SETTING – writing

The questionnaire below is comprised of 4 sections, viz. Standard Setting Session, Cambridge ESOL and Eurocentres Calibrated Samples, CEF Descriptors and CEF Linking Process, which correspond to aspects of the process undertaken as part of the writing standard setting workshop.

I would be grateful if you would indicate your level of agreement with each of the statements given and add any additional general comments you might have, or any specific comments about any statement in the questionnaire. You can use the back of these sheets, if necessary, to enlarge upon your comments. The information you provide will remain confidential.

I. Standard Setting Session

Item	Statement	Strongly Agree	Agree	Disagree	Strongly Disagree
1	Introductory talk provided me with a clear understanding of the purpose of the session.				
2	The tasks were clearly explained.				
3	The training and practice exercises helped me understand how to perform the tasks.				
4	Discussions aided my understanding of the levels.				
5	Discussions aided my understanding of how performances are assessed.				
6	There was adequate time provided for doing the tasks.				
7	There was adequate time provided for discussions.				
8	There was an equal opportunity for everyone to contribute his/her ideas and opinions.				
9	I was able to follow the instructions accurately.				
10	I was able to complete the judgment sheets accurately.				
11	The discussions after the first round of ratings were helpful to me.				
12	The discussions after the second round of ratings were helpful to me.				
13	I am confident about the defensibility of the final recommended cut score.				
14	I am confident about the appropriateness of the final recommended cut score.				
15	There was a productive and efficient working environment.				

Adapted from: Cizek, Bunch, & Koons (2004)

II. Cambridge ESOL Calibrated Samples

Item	Statements	Strongly Agree	Agree	Disagree	Strongly Disagree
1	The *written* samples helped me to carry out the standardization task.				
2	The *spoken* samples helped me to carry out the standardization task.				
3	The *written* samples were representative of the levels claimed by the exam providers.				

III. CEF Descriptors and Assessment Scales

Item	Statements	Strongly Agree	Agree	Disagree	Strongly Disagree
1	The Written Assessment Criteria Grid was easy to use.				
2	I felt the need to assign + levels for some written samples.				
3	The actual use of the CEF written descriptors showed me that they are applicable to *any* context.				
4	The actual use of the CEF written descriptors showed me that they are applicable to *academic* context.				
5	The CEF written descriptors catered for all aspects of writing skills.				

IV. CEF Linking Process

Item	Statements	Strongly Agree	Agree	Disagree	Strongly Disagree
1	The process forces BUSEL to consider the level of the COPE exam.				
2	The process is beneficial for BUSEL in seeing the actual level of the papers in the COPE exam.				
3	The process clearly pinpoints areas for revision/ reconsideration if BUSEL were to increase the standards of the COPE exam.				
4	The linking process makes COPE a more reliable exam.				
5	The linking process contributes to the validation of the COPE exam.				

Comments:

Thank you very much for the time you have put into answering these questions. Your answers will be collated and aggregated for use as part of my PhD research.

APPENDIX 3H Standardisation stage field notes coding scheme

Themes	Descriptions	Skill	Examples
Test taker	Experiential	W	They are used to writing academic essays FCE probably thinks the student has control over stereotypical language
Context validity	Task design	W	You can't penalize students who don't go beyond the task The task is certainly limiting the amount of production Out tasks should be parallel
	Task demands	W	You have to keep the reader in mind They have to expand in some length The student did not tackle many elements of the task
Cognitive validity	Language knowledge	W	It doesn't fit in with the high degree of grammatical control There is quite a broad range in this paper
	Content knowledge	W	The task itself is not complex, just family issues
Scoring validity	Criteria	W	There is an overlap between B2 and C1 regarding argument If we look at B2 and C2 what is the difference between the well-structured and the C1 descriptor?
	Prompt	W	Task one seems easier than task two
Implications of the level of the exam	Understanding of level	W	B2 is just an intermediate level but the upper end of B2 is a very high level A clear B2 is enough to study at Bilkent
Context validity	Task design	R	The text is authentic, you need background information to understand this

	Task demands	R	Items have a lot of low frequency words There are a lot of clues in the text to get the answer They can't use a dictionary in the exam
Cognitive validity	Language knowledge	R	A least able B2 candidate could not answer this question because it is related to the word rich species. If they don't know this, they won't get it. The attitude ones are a bit tricky because of the fine nuances of meaning
	Content knowledge	R	Background information is needed to get this question
Scoring validity	Criteria	R	Are these descriptors based on people who are immersed in the language Inferencing is borne out by the last two sentence so of our descriptor
	Items	R	The distracters are weak Selected response makes the items easy
Implications of the level of the exam	Intended level	R	This requires a lot interpretation and the least able B2 wouldn't be able to do that
	Understanding of level	R	We seem to be a band below Cambridge with our judgments Text and item you give one level and then if you take the strategies into consideration, your judgment might change
	Increasing the standards	R	Even the number of distracters go in the equation and makes the test easy

APPENDIX 3K Empirical validation stage interview coding scheme

Themes	Descriptions	Skill	Examples
Role of the empirical validation stage	Statistical information	R/W	It tells you where to get statistical information about your exam
	Validating the exam	R	We validated the reading exam by linking it to other exams
	Validating the link	R/W	But if we look beyond that you're empirically validating your exam and the procedures throughout from the very beginning of the project
Contribution of the empirical validation stage to the validity of COPE	Cognitive aspects	R	The cognitive requirements of the questions and how students respond to them, the type of questions we are asking are pretty similar to the other exams
	Design	R	When I look at the design of our reading, it is very similar design with all the problems with multiple choice
	Criterion-related aspects	R/W	It was a check that our understanding for COPE B2 level exam was pretty much in line with the Cambridge B2 level We have only teacher judgments … we can't take FCE because it is completely at odds with our construct
	Consequential aspect	R/W	We have seen that a large proportion of our students are not at B2 level
Adjusting the level of the exam	Understanding the level	R/W	Particularly been valuable in looking at the level we expect and realistically seeing the number of students who get there
	Adjusting	R/W	We have seen that some items are clearly below the level The level of some COPE texts is below the level. Those texts would now have to be looked at

APPENDIX 3L Phase 2 questionnaire

Dear Colleague,

I would like to invite you to take part in a small scale study which aims at investigating the contributions of the CEFR linking process to the COPE exam. The attached questionnaire is comprised of 7 sections, viz. test taker, context validity, cognitive validity, scoring validity, consequential validity, and criterion-related validity.

I would be grateful if you would fill in the questionnaire and add any additional general comments you might have, or any specific comments about any statement in the questionnaire. You can use the back of these sheets, if necessary, to enlarge upon comments. The information you provide will remain confidential and anonymity is assured. However, it is important to write does your names so that I can arrange interviews with some of you if necessary to clarify issues that may arise as a result of the questionnaire.

I would also like to note here that there might be stages of the CEFR project that you did not take part in such as the empirical validation stage in particular. If you feel you cannot make comments about those stages you can write NA in the relevant boxes.

I very much hope that you will feel able to participate. May I thank you, in advance, for your valuable cooperation.

Elif Kantarcıoğlu

Contact details (kutevu@bilkent.edu.tr and ext. 5244)

I. TEST TAKER

Which (if any) of the following test taker characteristics did you take into consideration and at what stage? You can tick more than one stage for each item. For each stage, there are two boxes: for writing and reading linking separately. Please tick the one that is relevant or both.

	STAGES OF THE CEFR LINKING PROCESS			
TEST TAKER CHARACTERISTICS	Familiarisation	Specification	Standardisation	Empirical Validation
1. Physical/physiological needs (e.g. Braille copies, enlarged print versions, etc.)	R□ W□	R□ W□	R□ W□	R□ W□
2. Psychological characteristics (e.g. learning styles, personality, emotional state, etc.)	R□ W□	R□ W□	R□ W□	R□ W□
3. Experiential characteristics (e.g. familiarity with the test)	R□ W□	R□ W□	R□ W□	R□ W□

II. CONTEXT VALIDITY

The appropriacy of which (if any) of the following areas in relation to the target context did you take into consideration and at what stage? You can tick more than one stage for each item. For each stage, there are two boxes: for writing and reading linking separately. Please tick the one that is relevant or both.

	STAGES OF THE CEFR LINKING PROCESS			
ELEMENTS OF CONTEXT VALIDITY	Familiarisation	Specification	Standardisation	Empirical Validation
1. Rubrics / prompts	R□ W□	R□ W□	R□ W□	R□ W□
2. Purpose of a task	R□ W□	R□ W□	R□ W□	R□ W□
3. Response format (e.g. short answer, MC, free response, etc.)	R□ W□	R□ W□	R□ W□	R□ W□
4. Marking criteria	R□ W□	R□ W□	R□ W□	R□ W□
5. Weighting of an item or section (points allocated)	R□ W□	R□ W□	R□ W□	R□ W□
6. Order of items	R□ W□	R□ W□	R□ W□	R□ W□
7. Time constraints	R□ W□	R□ W□	R□ W□	R□ W□
8. Discourse mode (genre, text type, etc.)	R□ W□	R□ W□	R□ W□	R□ W□
9. Channel of communication (use of graphs, charts, multiple tasks, etc.)	R□ W□	R□ W□	R□ W□	R□ W□
10. Text length	R□ W□	R□ W□	R□ W□	R□ W□
11. Nature of information in the text (abstract vs. concrete)	R□ W□	R□ W□	R□ W□	R□ W□
12. Content knowledge (topic	R□ W□	R□ W□	R□ W□	R□ W□

content)				
13. Lexical density in the input and output text	R□ W□	R□ W□	R□ W□	R□ W□
14. Structural density in the input and output text	R□ W□	R□ W□	R□ W□	R□ W□
15. Functional density in the input and output text (advise, persuade, describe, etc.)	R□ W□	R□ W□	R□ W□	R□ W□
16. Audience	R□ W□	R□ W□	R□ W□	R□ W□
17. Physical conditions of test administration	R□ W□	R□ W□	R□ W□	R□ W□
18. Uniformity of test administration	R□ W□	R□ W□	R□ W□	R□ W□
19. Security of the test	R□ W□	R□ W□	R□ W□	R□ W□

III. COGNITIVE VALIDITY

Which (if any) of the following areas did you take into consideration and at what stage? You can tick more than one stage for each item. Note that the first part is related to reading only whereas the second part is related writing.

	STAGES OF THE CEFR LINKING PROCESS			
READING	Familiarization	Specification	Standardization	Empirical Validation
1. Type of reading (careful, expeditious, etc.)	□	□	□	□
2. Sub-skills involved	□	□	□	□
3. Strategies involved	□	□	□	□
4. Purpose	□	□	□	□
5. Monitoring own reading	□	□	□	□
6. Word recognition	□	□	□	□
7. integration with the previous parts of the text	□	□	□	□
8. Grammatical knowledge	□	□	□	□
9. Textual knowledge (cohesion and coherence)	□	□	□	□
10. Functional (pragmatic) knowledge	□	□	□	□
11. Sociolinguistic knowledge	□	□	□	□
12. Background knowledge of the topic	□	□	□	□
13. Knowledge provided in the	□	□	□	□

text				
14. Appropriateness of the response format (MC, open ended, etc.)	□	□	□	□
WRITING	□	□	□	□
1. Type of writing (careful, expeditious, etc.)	□	□	□	□
2. Sub-skills involved	□	□	□	□
3. Strategies involved	□	□	□	□
4. Topic and genre modifying	□	□	□	□
5. Generating ideas	□	□	□	□
6. Organizing ideas	□	□	□	□
7. Putting ideas into appropriate, cohesive and coherent language	□	□	□	□
8. Evaluating and revising own writing	□	□	□	□
9. Grammatical knowledge	□	□	□	□
10. Textual knowledge (cohesion and coherence)	□	□	□	□
11. Functional (pragmatic) knowledge	□	□	□	□
12. Sociolinguistic knowledge	□	□	□	□
13. Background knowledge of the topic	□	□	□	□
14. Knowledge expected in the output text	□	□	□	□
15. The response format (MC, open ended, etc.)	□	□	□	□

IV. SCORING VALIDITY

Which (if any) of the following areas did you take into consideration and at what stage? You can tick more than one stage for each item. Note that the first part is related to reading only whereas the second part is related writing.

	STAGES OF THE CEFR LINKING PROCESS			
READING	Familiarisation	Specification	Standardisation	Empirical Validation
1. Item analysis	□	□	□	□
2. Internal consistency of the test (reliability)	□	□	□	□

3. Error of measurement	□	□	□	□
4. Marker reliability	□	□	□	□
5. Answer key	□	□	□	□
6. Training of markers/Standardization	□	□	□	□
7. Multiple marking	□	□	□	□
WRITING	□	□	□	□
1. Marking criteria	□	□	□	□
2. Holistic marking vs. analytical marking	□	□	□	□
3. Marker reliability	□	□	□	□
4. Marker consistency	□	□	□	□
5. Training of markers	□	□	□	□
6. Standardization	□	□	□	□
7. Multiple marking	□	□	□	□
8. Moderation of marking	□	□	□	□
9. Marking conditions	□	□	□	□
10. Grading and awarding	□	□	□	□

V. CONSEQUENTIAL VALIDITY

Which (if any) of the following areas did you take into consideration and at what stage? You can tick more than one stage for each item. For each stage, there are two boxes: for writing and reading linking separately. Please tick the one that is relevant or both.

	STAGES OF THE CEFR LINKING PROCESS			
ANALYSIS OF CONSEQUENTIAL VALIDITY	Familiarisation	Specification	Standardisation	Empirical Validation
1. Differential validity (analysis of bias)	R□ W□	R□ W□	R□ W□	R□ W□
2. Washback in classroom or workplace	R□ W□	R□ W□	R□ W□	R□ W□
3. Effect on individual within society	R□ W□	R□ W□	R□ W□	R□ W□

VI. CRITERION-RELATED VALIDITY

Which (if any) of the following areas did you take into consideration and at what stage? You can tick more than one stage for each item. For each stage, there are two boxes: for writing and reading linking separately. Please tick the one that is relevant or both.

	STAGES OF THE CEFR LINKING PROCESS			
ANLAYSIS OF CRITERION-RELATED VALIDITY	Familiarisation	Specification	Standardisation	Empirical Validation
1. Comparison with different versions of the same test	R□ W□	R□ W□	R□ W□	R□ W□
2. Comparison with the same test administered on different occasions	R□ W□	R□ W□	R□ W□	R□ W□
3. Comparison with other tests/measurements	R□ W□	R□ W□	R□ W□	R□ W□
4. Comparison with future performance	R□ W□	R□ W□	R□ W□	R□ W□

VII. INSTITUTIONAL IMPLICATIONS

Which (if any) of the following areas did the COPE CEFR linking project shed light on? You can tick more than one stage for each item. For each stage, there are two boxes: for writing and reading linking separately. Please tick the one that is relevant or both.

	STAGES OF THE CEFR LINKING PROCESS			
IMPLICATIONS	Familiarisation	Specification	Standardisation	Empirical Validation
1. Better understanding of what the COPE exam measures.	R□ W□	R□ W□	R□ W□	R□ W□
2. The level of the paper (Writing, reading).	R□ W□	R□ W□	R□ W□	R□ W□
3. Areas for revision	R□ W□	R□ W□	R□ W□	R□ W□
4. Areas that could be focused on to alter the level of the exam.	R□ W□	R□ W□	R□ W□	R□ W□
5. Do you think the level of the exam is suitable for its purpose? (For writing, pass-18 out of 30- is B2. For reading pass – 21 out of 35)	Yes □ No□			

APPENDIX 3N Phase 3 questionnaire

Dear Colleague,

The chart below aims to investigate what kind of validity evidence is available or gathered for the COPE examination. The first column entitled 'Validity Aspect' includes each aspect of validity with guiding statements. The second column deals with the reading paper and the third with the writing paper. Under the second and third columns you have three options to consider. For the given aspect of validity under column 1,

- if you think validity evidence has always been collected tick ALWAYS;
- if you think validity evidence has been collected as a result of the CEFR project tick CEFR;
- if you think no validity evidence is collected then tick NONE.

Please use the back of this page if you have any comments.

Thank you very much for your cooperation.

Elif Kantarcioglu

VALIDITY ASPECT	READING			WRITING		
	ALWAYS	CEFR	NONE	ALWAYS	CEFR	NONE
TEST TAKER CHARACTERISTICS						
How the physical/physiological characteristics of candidates are addressed by the test						
How the psychological characteristics of candidates are addressed by the test						
How the experiential characteristics of candidates are addressed by the test						
CONTEXT VALIDITY						
Whether the contextual characteristics of the test task are situationally fair to the candidates						
Whether the contextual characteristics of the test administration are situationally fair to the candidates						
COGNITIVE VALIDITY						
Whether the cognitive processes required to complete the tasks are interactionally authentic?						
SCORING VALIDITY						
How far we can depend on the scores of the test						
CONSEQUENTIAL VALIDITY						
What impact the test has on its various stakeholders						
CRITERION-RELATED VALIDITY						
What external evidence there is that the test is doing a good job						

APPENDIX 30 Consent form

Roehampton University

ETHICS BOARD

RESEARCH PARTICIPANT CONSENT FORM

Title and brief description of Research Project:

A Case-Study of the Process of Linking an Institutional English Language Proficiency Test (COPE) for Access to University Study in the Medium Of English to the Common European Framework for Languages: Learning, Teaching and Assessment

The aim of the research is to study and validate the process of linking the COPE exam to the CEFR. It also involves evaluating the Manual for relating language examinations to the Common European Framework of Reference for Languages: learning, teaching, assessment.

Name and status of Investigator:

Elif Kantarcıoğlu – PhD student at the University of Roehampton

Consent Statement:

I agree to take part in this research, and am aware that I am free to withdraw at any point. I understand that the information I provide will be treated in confidence by the investigator and that my identity will be protected in the publication of any findings.

Name …….

Signature ……

Date ……

Please note: if you have a concern about any aspect of your participation, please raise this with the investigator, or with the Director of Studies, who is Dr. Barry O'Sullivan.

Name: Dr. Barry O'Sullivan
Direct Phone No: 44 (0)20 8392 3348 Email: b.osullivan@roehampton.ac.uk

APPENDIX 4A Familiarisation stage questionnaire

CEF PROJECT FAMILIARISATION STAGE SESSION 1 QUESTIONNAIRE COLLATION

VI. Background Information

1. Years of experience as an English language teacher (Circle ONE)

0-4 / 1	5-9 / 4	10-14 / 5	15-19 / 1	20-24 / 1	25+

2. Role in BUSEL - Please tick one of the below as appropriate

Teacher 3 ☐
Head of Teaching Unit 1 ☐
Textbook Development Group Member ☐
Curriculum and Testing Unit Member 5 ☐
Member of the Directorate 1 ☐
Other (Please specify) _____Teacher trainer 1_____

3. How familiar were you with the CEF levels before you took part in the project? (Circle ONE)

Not familiar at all				Very familiar
1 / 3	2 / 3	3 / 4	4 / 2	5

4. Read the following statements and decide whether they are true or false. Put a tick under the right column. As you answer the questions, please keep in mind that all the statements are about your knowledge about the CEFR before you were involved in the project.

Before I was invited to take part in the project, I knew that	**FALSE**	**TRUE**
1. the CEFR is a document for language learning, teaching and assessment in Europe.		12
2. the CEFR consists of 6 levels.	11	1
3. exams such as FCE, IELTS or TOEFL are linked to the CEFR.	6	6

VII. Pre-session Reading Tasks

Circle ONE number for each of the statement below to give your opinion about the pre-session reading tasks.

	Disagree Strongly	Disagree	Not sure	Agree	Agree Strongly
1. The amount of the reading material was manageable.			2	5	5
2. The amount of time given for the reading was sufficient.			1	2	9
3. The aim of the tasks was clear.			2	5	5
4. The tasks had clear instructions.			1	3	8
5. The task sheets were well-designed.				5	7
6. The content of the reading material was easy to comprehend.		1	7	4	

7. Doing the tasks helped me see how the CEF levels progressed.			2	7	3

WEIGHTED TOTALS

	Disagree Strongly	Disagree	Not sure	Agree	Agree Strongly
1. The amount of the reading material was manageable.			6	20	25
2. The amount of time given for the reading was sufficient.			3	8	45
3. The aim of the tasks was clear.			6	20	25
4. The tasks had clear instructions.			3	12	40
5. The task sheets were well-designed.				20	35
6. The content of the reading material was easy to comprehend.		2	21	16	
7. Doing the tasks helped me see how the CEF levels progressed.			6	28	15

MEANS

	TOTAL	MEAN
1. The amount of the reading material was manageable.	51	4.25
2. The amount of time given for the reading was sufficient.	56	4.66
3. The aim of the tasks was clear.	51	4.25
4. The tasks had clear instructions.	55	4.58
5. The task sheets were well-designed.	55	4.58
6. The content of the reading material was easy to comprehend.	39	3.25
7. Doing the tasks helped me see how the CEF levels progressed.	49	4.08

VIII. Familiarization Session 1

Circle ONE number for each of the statements below to give your opinion about the familiarization session. (For questions 7 to 10, the tasks used in the session are explained at the end of the table.)

	Disagree Strongly	Disagree	Not sure	Agree	Agree Strongly
1. The purpose of the session was clear.				9	3
2. The opening speech about the CEF project gave me a clear idea about the project.			4	5	3
3. The opening speech about the CEF project gave me a clear idea of my role in the project.			6	5	1
4. The session was well-designed.				8	4
5. There was a clear progression of tasks in the session.				8	3
6. The pre-session reading tasks were linked to the tasks in the session itself.			2	5	5

7. Task 1, 2 and 3 followed by a discussion helped clarify the CEF levels better.			2	6	4
8. Task 4 helped me better understand the CEF levels.			4	6	1
9. At the end of task 5 and 6, I began to see the differences between the CEF levels.			3	6	3

The tasks carried out in the session are as follows:

Task 1: Identifying key characteristics of the levels in the CEF global scale
Task 2: Analyzing section 3.6 – salient features of CEF levels
Task 3: Poster completion
Task 4: Relating BUSEL levels to CEF levels
Task 5: Reordering CEF global scales followed by feedback and discussion
Task 6: Reordering CEF scales for each skill followed by feedback and discussion

WEIGHTED TOTALS

	Disagree Strongly	Disagree	Not sure	Agree	Agree Strongly
1. The purpose of the session was clear.				36	15
2. The opening speech about the CEF project gave me a clear idea about the project.			12	20	15
3. The opening speech about the CEF project gave me a clear idea of my role in the project.			18	20	5
4. The session was well-designed.				32	20
5. There was a clear progression of tasks in the session.				32	15
6. The pre-session reading tasks were linked to the tasks in the session itself.			6	20	25
7. Task 1, 2 and 3 followed by a discussion helped clarify the CEF levels better.			6	24	20
8. Task 4 helped me better understand the CEF levels.			12	24	5
9. At the end of task 5 and 6, I began to see the differences between the CEF levels.			9	24	15

MEANS

	TOTALS	MEAN
1. The purpose of the session was clear.	51	4.25
2. The opening speech about the CEF project gave me a clear idea about the project.	47	3.91
3. The opening speech about the CEF project gave me a clear idea of my role in the project.	43	3.58
4. The session was well-designed.	52	4.33
5. There was a clear progression of tasks in the session.	47	3.91
6. The pre-session reading tasks were linked to the tasks in the session itself.	51	4.25
7. Task 1, 2 and 3 followed by a discussion helped clarify the	50	4.16

CEF levels better.		
8. Task 4 helped me better understand the CEF levels.	41	3.41
9. At the end of task 5 and 6, I began to see the differences between the CEF levels.	48	4.00

Content

Circle ONE number for each of the statements below to give your opinion about the content of familiarization session 1.

	Disagree Strongly	Disagree	Not sure	Agree	Agree Strongly
1. There is a clear progression between the CEF global scales (CEF Table 1).			7	5	
2. There is a clear progression between the CEF self-assessment scales for each skill (CEF Table 2).			5	7	
3. The number of levels in CEF is adequate to show language progression.		3	4	4	1
4. CEF levels can be used in every context.		2	9	1	
5. The can-do statements used to describe the levels are unambiguous.	1	1	5	5	
6. I can easily relate the CEF levels to our context.		2	9	1	
7. There is a clear link between CEF levels and the levels commonly used by commercial books i.e. Elementary, Intermediate, Advanced etc.		3	6	3	
8. Aspects of language use are clearly explained in Chapter 4 of the CEF booklet.		2	5	5	
9. Language competences are clearly explained in Chapter 5 of the CEF booklet.		2	5	5	
10. I can now understand the rationale behind the CEF descriptor.		2	4	4	2

WEIGHTED TOTALS

	Disagree Strongly	Disagree	Not sure	Agree	Agree Strongly
1. There is a clear progression between the CEF global scales (CEF Table 1).			21	20	
2. There is a clear progression between the CEF self-assessment scales for each skill (CEF Table 2).			15	28	
3. The number of levels in CEF is adequate to show language progression.		6	12	16	5
4. CEF levels can be used in every context.		4	27	4	
5. The can-do statements used to describe the levels are unambiguous.	1	2	15	20	
6. I can easily relate the CEF levels to our context.		4	27	4	
7. There is a clear link between CEF levels and the levels commonly used by commercial books i.e. Elementary,		6	18	12	

Intermediate, Advanced etc.					
8. Aspects of language use are clearly explained in Chapter 4 of the CEF booklet.		4	15	20	
9. Language competences are clearly explained in Chapter 5 of the CEF booklet.		4	15	20	
10. I can now understand the rationale behind the CEF descriptor.	2	4	12	16	

MEANS

	TOTALS	MEAN
1. There is a clear progression between the CEF global scales (CEF Table 1).	41	3.41
2. There is a clear progression between the CEF self-assessment scales for each skill (CEF Table 2).	43	3.58
3. The number of levels in CEF is adequate to show language progression.	39	3.25
4. CEF levels can be used in every context.	35	2.91
5. The can-do statements used to describe the levels are unambiguous.	38	3.16
6. I can easily relate the CEF levels to our context.	35	2.91
7. There is a clear link between CEF levels and the levels commonly used by commercial books i.e. Elementary, Intermediate, Advanced etc.	36	3.00
8. Aspects of language use are clearly explained in Chapter 4 of the CEF booklet.	39	3.25
9. Language competences are clearly explained in Chapter 5 of the CEF booklet.	39	3.25
10. I can now understand the rationale behind the CEF descriptor.	34	2.83

IX. Future needs / demands

As you know, familiarization sessions will continue. Please write down any suggestions or things you would like to see done/covered in the follow-up sessions.

- *We need to start questioning the descriptors at more detail.*
- *We need to work with the descriptors using real samples – might help us more*
- *When you send us reading tasks, could you please indicate which chapters to focus on more?*

APPENDIX 4B Outline of familiarisation session 1 and session notes

CEF LINKING PROJECT
STAGE 1: FAMILIARIZATION
SESSION 1

DATE: 03-04.07.06
PLACE: DB 01

PRIOR TO THE SESSION

- As the group is almost totally unfamiliar with the CEF levels, every group member needs to spend some time analyzing some of the CEF documents (Chapters 3, 4 and 5). The members will receive tasks to go with the CEF pack to help them with familiarization. *Preferably about 1 or 2 months in advance, the packs need to be sent out.*(See Appendix 1 for the details about the pack)

- The group members will be asked to send any questions or comments as they read/analyze the documents. These will be collated to be discussed at the familiarization session.

- The process to be followed in analyzing the judgment agreement needs to be drawn. For each analysis, the level descriptors will be labeled so that the judgments can be analyzed in FACETS (Multi-facet Rasch).

- Report sheets will be prepared to be used in sharing the results of the Multi-facet Rasch analysis with the judges. Personal report sheets and an overall report format for general discussion will be prepared.

- A questionnaire will be used to evaluate the effectiveness of the familiarization session.

- Materials need to be prepared for the session a week in advance to the session.

SESSION PLAN

DAY 1

Step 1 – Introduction by JOD followed by the session leader(s) (30 mins)
- purpose of the study (CT)
- background to CEF (CT)
- overview of the project (EKAN)
- overview of the training (EKAN)

Step 2 – A General Analysis of CEF levels in relation to BUSEL levels (CT)

AIM: *to form a group consensus as to what the key elements / indicators of CEF levels are and relate the CEF to the BUSEL levels.*

- groups look back at the pre-session tasks, go over them again and identify the features / key words of a certain level considering what makes a level different from the other (write key words on posters and tick the ones they agree) (10 mins)
- participants are given CEF section 3.6 (CEF level specifications) to raise awareness of the salient features of each of the CEF levels (15 mins)
- groups go back to the posters prepared and add further features
- 6 volunteers take the posters off the walls and go over them for the group

Coffee Break

- participants carry out a discussion on which of the CEF levels would be more suitable for a COPE level student (keeping in mind the key words, focusing on the current situation and the ideal)
- participants carry out a discussion on which BUSEL levels correspond to which CEF levels (start with UPP and go down)

Step 3 – Using the CEF global scale to produce a ranked list of can-do statements and compare with the original (EKAN) (15 mins)

AIM: *to investigate how the participants perceive the progression of the CEF the levels*

- each participant is given an envelope with a set of can-do statements for each CEF level (6 levels)
- explain why the tasks will be carried out individually (everyone is being trained to be experts and everyone's individual input, perception is important)
- participants individually sort the can-do statements into levels based on their own perceptions and fill in the chart given (10 mins)

LUNCH BREAK

- during the lunch break the data from the charts is analyzed using MFR
- after the break the participants are given feedback on how they've done and their answers are compared with the original CEF global scale
- in cases where the participants' perceptions are different or do not reflect the original CEF scale, participants discuss possible reasons

Step 4 – Using the Self-assessment grid to produce a ranked list of can-do statements for each skill and compare with the original (EKAN)

AIM: *to investigate how the participants perceive the progression of the CEF the levels on a skills basis*

- participants are provided with a template on A3 paper and an envelope with a set of can-do statements for each skill and CEF level (6 levels) (guide participants – sort out the skills first and then grade)

- participants individually reconstruct the CEF Table 2 – self-assessment grid based on their own perceptions
- discussion on how the participants felt as they were during the task

END OF DAY 1

- at the end of DAY 1 the data form the templates is analyzed using MFR – decide on how to give feedback

DAY 2

- the participants are given feedback on how they've done and their answers are compared with the original CEF Table 2
- in cases where the participants' perceptions are different or do not reflect the original CEF self-assessment grid, participants discuss possible reasons

Step 5 – Comparison of the CEF global scale and the self-assessment grid (CT) (photocopy pages 24, 26 and 27)

AIM: *to further analyze and compare the CEF global scale and the self-assessment grid*

- the participants will be asked to compare the CEF global scale Table 1 and the self-assessment grid Table 2 (15 mins) – anything that is missing in the global scale but mentioned in the self-assessment grid
- any mismatches identified will be documented and discussed

Step 6 – Wrap-up

- A list of questions or issues raised either by the session leader(s) or the participants will be put up to be discussed as a group (content-specific questions or comments sent by the participants prior to the session)
- Time is allocated for participants to raise any other issues.
- The points are discussed one by one

Step 7 – Evaluation

- participants are given a questionnaire to reflect on the familiarization stage
- the data is collated and analyzed
- a follow up session will be designed based on the MFR and questionnaire results of Session 1

Data collection will be done as follows:

1. The session will be video recorded.
2. Researcher will keep field notes.
3. Follow-up interviews may be carried out.

APPENDIX 4C Specification stage group interview questions

1. What is the role of the specification stage in the CEFR linking process?

2. Did the process of filling in the specification forms contributed to your understanding or knowledge of the cope exam, reading and writing papers in particular? Think in terms of
    - design
    - cognitive requirements
    - scoring
    - score interpretation
    - score value.

3. Did the process of filling in the specification forms tell you anything about the institutional standards set through the cope exam reading and writing papers?

4. If you were to increase the level of the cope exam, do you think the specification stage would be of any help to you?

APPENDIX 4D Standardisation stage writing questionnaire results

EVALUATION OF COPE STANDARD SETTING COLLATED DATA - writing

I. Standard Setting Session

Table 1: Distribution of responses to each question

Question	Strongly Agree	Agree	Disagree	Strongly Disagree	Missing Data	Total
1	6	6				12
2	6	6				12
3	8	4				12
4	10	2				12
5	8	4				12
6	7	5				12
7	7	5				12
8	7	4	1			12
9	10	2				12
10	9	3				12
11	11	1				12
12	6	6				12
13	6	6				12
14	7	5				12
15	10	2				12

Table 2: Frequencies of responses grouped into two categories

Question	Strongly Agree / Agree	Disagree / Strongly Disagree
1	12	
2	12	
3	12	
4	12	
5	12	
6	12	
7	12	
8	11	1
9	12	
10	12	
11	12	
12	12	
13	12	
14	12	
15	12	

Table 3: Proportions of responses in two categories

Question	Strongly Agree / Agree	Disagree / Strongly Disagree	%
1	100		100
2	100		100
3	100		100
4	100		100
5	100		100
6	100		100
7	100		100
8	91,66	8,33	100
9	100		100
10	100		100
11	100		100
12	100		100
13	100		100
14	100		100
15	100		100

Table 4: Weighted totals for each question

Question	Strongly Agree	Agree	Disagree	Strongly Disagree	Total	Mean
1	24	18			42	3,5
2	24	18			42	3,5
3	32	12			44	3,66
4	40	6			46	3,83
5	32	12			44	3,66
6	28	15			43	3,58
7	28	15			43	3,58
8	28	12	2		42	3,5
9	40	6			46	3,83
10	36	9			45	3,75
11	44	3			47	3,91
12	24	18			42	3,5
13	24	18			42	3,5
14	28	15			43	3,58
15	40	6			46	3,83

II. Cambridge ESOL and IELTS Samples

Table 1: Distribution of responses to each question

Question	Strongly Agree	Agree	Disagree	Strongly Disagree	Missing Data	Total
1	7	5				12
2	4	3	5		1**	12
3	3	3	4	1	1**	12

- 1 person didn't attend the speaking paper standard setting session and thus didn't answer these questions.

Table 2: Frequencies of responses grouped into two categories

Question	Strongly Agree / Agree	Disagree / Strongly Disagree	Total
1	12		12
2	7	5	12
3	6	5	11

Table 3: Proportions of responses in two categories

Question	Strongly Agree / Agree	Disagree / Strongly Disagree	%
1	100		100
2	58,3	41,6	100
3	54,54	45,45	100

Table 4: Weighted totals for each question

Question	Strongly Agree	Agree	Disagree	Strongly Disagree	Total	Mean
1	28	15			43	3,58
2	16	9	10		35	2,91
3	12	9	8	1	30	2,72

III. CEF Descriptors and Assessment Scales

Table 1: Distribution of responses to each question

Question	Strongly Agree	Agree	Disagree	Strongly Disagree	Missing Data	Total
1	2	6	4			12
2	4	7	1			12
3		9	3			12
4	1	8	2	1		12
5	1	5	5		1*	12

- 1 person ticked on the line between agree and disagree

Table 2: Frequencies of responses grouped into two categories

Question	Strongly Agree / Agree	Disagree / Strongly Disagree	Total
1	8	4	12
2	11	1	12
3	9	3	12
4	9	3	12
5	6	5	11

Table 3: Proportions of responses in two categories

Question	Strongly Agree / Agree	Disagree / Strongly Disagree	%
1	66,6	33,3	100
2	91,6	8,3	100
3	75	25	100
4	75	25	100
5	54,54	45,45	100

Table 4: Weighted totals for each question

Question	Strongly Agree	Agree	Disagree	Strongly Disagree	Total	Mean
1	8	18	8		34	2,8
2	16	21	2		39	3,25
3		27	6		33	2,75
4	4	24	4	2	34	2,8
5	4	15	10		29	2,63

IV. CEF Linking Process

Table 1: Distribution of responses to each question

Question	Strongly Agree	Agree	Disagree	Strongly Disagree	Missing Data	Total
1	5	6	1			12
2	8	4				12
3	4	8				12
4	5	6		1		12
5	9	3				12

Table 2: Frequencies of responses grouped into two categories

Question	Strongly Agree / Agree	Disagree / Strongly Disagree	Total
1	11	1	12
2	12		12
3	12		12
4	11	1	12
5	12		12

Table 3: Proportions of responses in two categories

Question	Strongly Agree / Agree	Disagree / Strongly Disagree	%
1	91,6	8,3	100
2	100		100
3	100		100
4	91,6	8,3	100
5	100		100

Table 4: Weighted totals for each question

Question	Strongly Agree	Agree	Disagree	Strongly Disagree	Total	Mean
1	20	18	2		40	3,3
2	32	12			44	3,6
3	16	24			40	3,3
4	20	18		1	39	3,25
5	36	9			45	3,75

COMMENTS

I. Standard Setting Session

Item 6 – Sometimes too much time was given

II. Cambridge ESOL Calibrated Samples

Items 1 & 2 – quite different but allows to get familiar with criteria
Item 3 – seemed a bit high (1 level)
Item 3 – obviously half or one CEFR band lower

III. CEF Descriptors and Assessment Scales

Item 7 – some boxes not described very well i.e. Range B2
Items 9 & 10 – but more in some cases
Item 6 – maybe not content
Items 1, 2 & 3 – we discussed this issue, they were designed for use in different contexts
Items 4 & 5, 8 & 9 – this is a tricky question. If adapted properly they might be used

IV. CEF Linking Process

Items 6 & 7 – a linking process cannot do this on its own but contribute

- Overall it was a very productive three days. I feel that we are now well on track to a successful end to this process.
- Thank you very much
- Thanks a lot for everything, Elif. It was a great opportunity for me to learn lots of things about CEF and COPE levels.

APPENDIX 4E Standardisation stage reading questionnaire results

EVALUATION OF COPE STANDARD SETTING COLLATED DATA – READING

I. Standard Setting Session

Table 1: Distribution of responses to each question

Question	Strongly Agree	Agree	Disagree	Strongly Disagree	Missing Data	Total
1	5	7				12
2	3	8	1			12
3	3	9				12
4	2	8	2			12
5	2	7	2		1	12
6	2	5	5			12
7	2	2	8			12
8	3	7	1	1		12
9	5	6	1			12
10	2	9	1			12
11	2	6	3	1		12
16	9	3				12

Table 2: Frequencies of responses grouped into two categories

Question	Strongly Agree / Agree	Disagree / Strongly Disagree	Missing Data
1	12		
2	11	1	
3	12		
4	10	2	
5	9	2	1
6	7	5	
7	4	8	
8	10	2	
9	11	1	
10	11	1	
11	8	4	
16	12		

Table 3: Proportions of responses in two categories

Question	Strongly Agree / Agree	Disagree / Strongly Disagree	%
1	100		100
2	91,66	8,33	100
3	100		100
4	83,33	16,66	100
5	75	16,66	100
6	58,33	41,66	100
7	33,33	66,66	100
8	83,33	16,66	100
9	91.66	8,33	100
10	91,66	8,33	100
11	66,66	33,33	100
16	100		100

Table 4: Weighted totals for each question

Question	Strongly Agree	Agree	Disagree	Strongly Disagree	Total	Mean
1	20	21			41	3,41
2	12	24	2		38	3,16
3	12	27			39	3,25
4	8	24	4		36	3
5	8	21	4		33	3*
6	8	15	10		33	2,75
7	8	3	16		27	2,25
8	12	21	2	1	36	3
9	20	18	2		40	3,33
10	8	27	2		37	3,08
11	8	18	6	1	33	2,75
16	36	9			45	3,75

* 1 person didn't answer this question

II. Cambridge ESOL and Finnish Matriculation Exams

Table 1: Distribution of responses to each question

Question	Strongly Agree	Agree	Disagree	Strongly Disagree	Missing Data	Total
1		9	2	1		12
2		10	1	1		12
3	1	11				12
4	1	2	7	2		12
5	1	6	3	2		12
6		6	6			12
7		1	7	4		12
8		6	2	4		12
9		9	1	1	1	12

Table 2: Frequencies of responses grouped into two categories

Question	Strongly Agree / Agree	Disagree / Strongly Disagree	Total
1	9	3	12
2	10	2	12
3	12		12
4	3	9	12
5	7	5	12
6	6	6	12
7	1	11	12
8	6	6	12
9	9	2	11

Table 3: Proportions of responses in two categories

Question	Strongly Agree / Agree	Disagree / Strongly Disagree	%
1	75	25	100
2	83,33	16,66	100
3	100		100
4	25	75	100
5	58,33	41,66	100
6	50	50	100
7	8,33	91,66	100
8	50	50	100
9	81,81	18,18	100

Table 4: Weighted totals for each question

Question	Strongly Agree	Agree	Disagree	Strongly Disagree	Total	Mean
1		27	4	1	32	2,66
2		30	2	1	33	2,75
3	4	33			37	3,08
4	4	6	14	2	26	21,66
5	4	18	6	2	30	2,5
6		18	12		30	2,5
7		3	14	4	21	1,75
8		18	4	4	26	21,66
9		27	2	1	30	2,72*

* 1 person didn't answer this question

III. CEF Descriptors

Table 1: Distribution of responses to each question

Question	Strongly Agree	Agree	Disagree	Strongly Disagree	Missing Data	Total
1	6	5	1			12
2	1	6	4	1		12
3	1	5	5	1		12
4		3	6	3		12

Table 2: Frequencies of responses grouped into two categories

Question	Strongly Agree / Agree	Disagree / Strongly Disagree	Total
1	11	1	12
2	7	5	12
3	6	6	12
4	3	9	12

Table 3: Proportions of responses in two categories

Question	Strongly Agree / Agree	Disagree / Strongly Disagree	%
1	91,66	8,33	100
2	58,33	41,66	100
3	50	50	100
4	25	75	100

Table 4: Weighted totals for each question

Question	Strongly Agree	Agree	Disagree	Strongly Disagree	Total	Mean
1	24	15	2		41	3,41
2	4	18	8	1	31	2,58
3	4	15	10	1	30	2,5
4		9	12	3	24	2

APPENDIX 5A Phase 2 questionnaire results

PART I

Which (if any) of the following **test taker characteristics** did you take into consideration and at what stage? You can tick more than one stage for each item. For each stage, there are two boxes: for writing and reading linking separately. Please tick the one that is relevant or both.

	STAGES OF THE CEFR LINKING PROCESS				
	Familiarization	Specification	Standardization	Empirical Validation	NA
1. Physical/physiological needs (e.g. Braille copies, enlarged print versions, etc.)	R□ W□	R1 W2	R1 W1	R□ W□	8
2. Psychological characteristics (e.g. learning styles, personality, emotional state, etc.)	R1 W□	R2 W1	R3 W2	R□ W□	7
3. Experiential characteristics (e.g. familiarity with the test)	R1 W1	R2 W2	R2 W2	R□ W□	8

PART II

Which (if any) of the following areas in relation to their **appropriacy to the target context** did you take into consideration and at what stage? You can tick more than one stage for each item. For each stage, there are two boxes: for writing and reading linking separately. Please tick the one that is relevant or both.

	STAGES OF THE CEFR LINKING PROCESS				
	Familiarization	Specification	Standardization	Empirical Validation	NA
1. Rubrics / prompts	R□ W□	R3 W3	R9 W9	R□ W□	1
2. Purpose of a task	R1 W2	R4 W4	R8 W7	R□ W□	2
3. Response format (e.g. short answer, MC, free response, etc.)	R1 W1	R5 W4	R9 W7	R□ W□	1
4. Marking criteria	R□ W1	R1 W2	R5 W8	R□ W□	2
5. Weighting of an item or section (points allocated)	R□ W□	R□ W□	R4 W2	R□ W□	6
6. Order of items	R□ W□	R1 W□	R5 W1	R□ W□	5
7. Time constraints	R□ W□	R2 W2	R7 W5	R□ W□	3
8. Discourse mode (genre, text type, etc.)	R□ W□	R4 W3	R10 W8	R□ W□	-
9. Channel of communication (use of graphs, charts, multiple tasks, etc.)	R□ W□	R□ W□	R1 W1	R□ W□	9
10. Text length	R1 W1	R3 W2	R10 W6	R□ W□	-
11. Nature of information in the text (abstract vs. concrete)	R3 W2	R3 W1	R10 W7	R□ W□	-
12. Content knowledge (topic	R1 W□	R3 W2	R7 W6	R□ W□	-

content)					
13. Lexical density in the input and output text	R2 W2	R4 W3	R9 W7	R□ W□	-
14. Structural density in the input and output text	R1 W1	R3 W3	R9 W6	R□ W□	-
15. Functional language in the input and output text (advise, persuade, describe, etc.)	R1 W1	R3 W3	R8 W5	R□ W□	2
16. Audience	R□ W□	R1 W2	R4 W4	R□ W□	6
17. Physical conditions of test administration	R□ W□	R1 W1	R2 W2	R□ W□	
18. Uniformity of test administration	R□ W□	R1 W1	R1 W1	R□ W□	
19. Security of the test	R□ W□	R1 W1	R1 W1	R□ W□	

PART III

Which (if any) of the following areas did you take into consideration and at what stage? You can tick more than one stage for each item. Note that the first part is related to reading only whereas the second part is related writing.

	STAGES OF THE CEFR LINKING PROCESS				
READING	Familiarization	Specification	Standardization	Empirical Validation	NA
1. Type of reading (careful, expeditious, etc.)	5	7	8	□	2
2. Sub-skills involved	4	6	9	□	1
3. Strategies involved	2	3	7	□	3
4. Purpose	4	7	6	□	3
5. Monitoring own reading (rereading, checking)	1	4	5	□	5
6. Word recognition	2	3	8	□	2
7. integration with the previous parts of the text	1	3	6	□	4
8. Grammatical knowledge	3	7	8	□	2
9. Textual knowledge (cohesion and coherence)	3	5	8	□	2
10. Functional knowledge	3	5	6	□	4
11. Sociolinguistic knowledge	3	5	7	□	3
12. Background knowledge of the topic	3	6	8	□	2
13. Knowledge provided in the	3	6	7	□	3

text					
14. Appropriateness of the response format (MC, open ended, etc.)	3	4	7	□	3
WRITING					
1. Type of writing (careful, expeditious, etc.)	4	6	6	□	4
2. Sub-skills involved	3	5	5	□	5
3. Strategies involved	2	3	4	□	6
4. Topic and genre modifying	5	5	7	□	3
5. Generating ideas	2	3	4	□	6
6. Organizing ideas	4	5	7	□	3
7. Putting ideas into appropriate, cohesive and coherent language	5	7	8	□	2
8. Evaluating and revising own writing	□	2	3	□	7
9. Grammatical knowledge	5	3	8	□	2
10. Textual knowledge (cohesion and coherence)	3	4	8	□	2
11. Functional knowledge	3	3	6	□	4
12. Sociolinguistic knowledge	3	3	6	□	4
13. Background knowledge of the topic	1	2	4	□	6
14. Knowledge expected in the output text	3	3	6	□	4
15. The response format (MC, open ended, etc.)	3	3	5	□	5

PART IV

Which (if any) of the following areas did you take into consideration and at what stage? You can tick more than one stage for each item. Note that the first part is related to reading only whereas the second part is related writing.

	STAGES OF THE CEFR LINKING PROCESS				
READING	Familiarization	Specification	Standardization	Empirical Validation	NA
1. Item analysis	1	1	7	2	2
2. Reliability of the test	1	2	6	2	4
3. Error of measurement in the test	-	-	1	2	8

4. Marker reliability	1	-	3	2	7
5. Answer key	1	1	2	-	8
6. Training of markers/Standardization	2	-	3	-	7
7. Multiple marking	-	-	2	-	8
WRITING					
1. Marking criteria	1	1	5	1	5
2. Holistic marking vs. analytical marking	1	2	5 □	-	-
3. Marker reliability	2	1	3	2	7
4. Marker consistency	1	1	3	2	7
5. Training of markers	1	-	2	2	8
6. Standardization	1	-	3	2	8
7. Multiple marking	-	-	1	1	9
8. Moderation of marking	-	-	1	-	9
9. Marking conditions	-	-	1	-	9
10. Grading and awarding	-	-	1	1	9

PART V

Which (if any) of the following areas did you take into consideration and at what stage? You can tick more than one stage for each item. For each stage, there are two boxes: for writing and reading linking separately. Please tick the one that is relevant or both.

	STAGES OF THE CEFR LINKING PROCESS				
	Familiarization	Specification	Standardization	Empirical Validation	NA
1. Differential validity (analysis of bias)	R□ W□	R□ W□	R1 W2	R□ W□	8
2. Washback in classroom or workplace	R□ W□	R□ W□	R2 W2	R□ W□	8
3. Effect on individual within society	R□ W□	R□ W□	R□ W□	R□ W□	10

PART VI

Which (if any) of the following areas did you take into consideration and at what stage? You can tick more than one stage for each item. For each stage, there are two boxes: for writing and reading linking separately. Please tick the one that is relevant or both.

	STAGES OF THE CEFR LINKING PROCESS				
	Familiarization	Specification	Standardization	Empirical Validation	NA

1. Comparison with different versions of the same test	R□ W□	R2 W	R4 W2	R□ W□	6
2. Comparison with the same test administered on different occasions	R□ W□	R□ W□	R□ W□	R□ W□	10
3. Comparison with other tests/measurements	R□ W□	R W	R3 W3	R1 W□	7
4. Comparison with future performance	R□ W□	R□ W□	R□ W□	R□ W□	10

PART VII

Which (if any) of the following areas did the **COPE CEFR linking project shed light on**? You can tick more than one stage for each item. For each stage, there are two boxes: for writing and reading linking separately. Please tick the one that is relevant or both.

	STAGES OF THE CEFR LINKING PROCESS				
	Familiarization	Specification	Standardizatio n	Empirical Validation	N A
1. Better understanding of what the COPE exam measures.	R1 W□	R4 W2	R9 W6	R1 W□	1
2. The level of the paper (Writing, reading).	R1 W2	R4 W3	R10 W8	R□ W□	-
3. Areas for revision	R□ W□	R5 W5	R9 W8	R□ W□	1
4. Areas that could be focused on to alter the level of the exam.	R□ W□	R2 W1	R9 W7	R□ W□	1
5. Do you think the level of the exam is suitable for its purpose? (For writing, pass-18 out of 30- is B2. For reading pass – 21 out of 35)	Reading Yes 9 No 1 Writing Yes 7 No 3				

APPENDIX 6A Phase 3 interview coding scheme

I: Interviewee

Fam: Familiarization

Spec: Specification

Stan: Standardization

EV: Empirical Validation

Themes	Descriptions	CEFR Stage	Interviewee 1	Interviewee 2	Interviewee 3
Validity of the CEFR linking process	Using the CEFR	Fam	We did the more extended familiarization that involved more sessions Enough background knowledge to be able to interpret it	We couldn't have gone forward if we stayed with the suggestions in the manual	If I couldn't understand what the CEFR was you know thoroughly, it would probably affect the judgments
	Forming connections	Fam	We also needed to know the background, and we also needed to know how to apply the CEFR in our own context	People were saying it's really difficult to relate the CEFR to the BUSEL to the academic context	No cases of this code for I3
	Confidence	Fam	No cases of this code for I1	There was a lot of uncertainty. People didn't confident	The quizzes that were given to check your understanding of the scales .. through the key questions they asked they raised my awareness of CEFR scales
	Seriousness	Fam	No cases of this code for I1	Partly I think the statistics added to the seriousness so people knew that they were	You the stats sort of shook people a little bit and they were trying to be more at the level

				being judged .. so people did their best	
	Group effort	Spec	Filling it all in to me is not as valid as sitting working on it, getting some, looking at what people are filling it out, what people have filled out and developing it further	It was an opportunity for the whole group to work closely with the descriptors Everybody questioned certain aspects and we all got a deeper understanding of COPE	Those discussions helped me get clearer Through those discussions in filling in the forms I think I did a better job
	Better understanding of COPE	Spec	Looking at what people filled out, working on it together was a better way of doing it	we all got a deeper understanding of COPE	No cases of this code for I3
	Cut scores	Stan	Using more than one standard setting method … makes the process more valid because you have more than one method and you can compare … the more data you get form different methods the healthier your cut score establishment becomes	It was very useful to look at different methods	No cases of this code for I3
	Advanced statistics	Stan	Because you have more than one method you can actually compare the stats	The more advanced statistical analysis that was that greatly contributed to the process	Maybe my judgments weren't reliable by doing these in a repeated way we got rid of level differences

	Confidence	Stan	The more experience raters have going through the process the more confident they become of their own judgments	People were pretty confident they sort of internalized the B2 level	
	Institutional bias	EV	We did send our papers out to external people to get their judgments … sometime over-familiarity with your context … can maybe blind you to certain things	COPE writing samples sent them to external people … that was very interesting. Again contributed to the validity	No cases of this code for I3
	Teacher judgments	EV	It's like another way of triangulating if you will getting extra data source to confirm your findings	Became part of the institutional culture … teacher shave become more aware of the B2 level and COPE	No cases of this code for I3
	External exams	EV	No cases of this code for I1	It helped to confirm that we do have a good understanding and we can write an exam at B2 level	We were given different papers from different exams and item were compared and it was useful
Validity of COPE	Using the CEFR	Fam	No cases of this code for I1	No cases of this code for I2	It was very useful to learn about the background to the CEFR and how the scales were developed as well as also this contributed to the validity of

					the COPE exam because if we couldn't understand the CEFR
	Assigning levels	Spec	No cases of this code for I1	No cases of this code for I2	That's more valid than just two people setting the level
	Cut scores	Stan	Using more than one standard setting method … makes the process more valid because you have more than one method and you can compare … the more data you get form different methods the healthier your cut score establishment becomes	It was very useful to look at different methods	No cases of this code for I3
	Advanced statistics	Stan	Because you have more than one method you can actually compare the stats	The more advanced statistical analysis that was that greatly contributed to the process	Maybe my judgments weren't reliable by doing these in a repeated way we got rid of level differences
	Confidence	Stan	The more experience raters have going through the process the more confident they become of their own judgments	People were pretty confident they sort of internalized the B2 level	No cases of this code for I3
	External people	EV	They are a little bit more objective than you are perhaps		No cases of this code for I3
	Expectations	EV	I think it gives you a much healthier picture of your	That increased our understanding	No cases of this code for I3

			students ability	of the level of the COPE and the kind of items were going to have in COPE	
	Evidence	EV	It's like another way of triangulating if you will getting extra data source to confirm your findings	We are adding to the validity evidence that we are gathering and that's reassuring	No cases of this code for I3